XI ХАЛЫҚАРАЛЫҚ ҒЫЛЫМИ-ТӘЖІРИБЕЛІК КОНФЕРЕНЦИЯ

«ИНФОРМАТИКА ЖӘНЕ ҚОЛДАНБАЛЫ МАТЕМАТИКА»

XI МЕЖДУНАРОДНАЯ НАУЧНО-ПРАКТИЧЕСКАЯ КОНФЕРЕНЦИЯ

«ИНФОРМАТИКА И ПРИКЛАДНАЯ МАТЕМАТИКА»

XI INTERNATIONAL SCIENTIFIC AND PRACTICAL CONFERENCE

«COMPUTER SCIENCE AND APPLIED MATHEMATICS»

CSAM 2026

МАТЕРИАЛЫ

КОНФЕРЕНЦИИ

Конференция материалдары • Conference Proceedings

ТОМ 2

Казахский национальный университет имени аль-Фараби
Институт информационных и вычислительных технологий КН МНВО РК
Люблинский технический университет (г. Люблин, Польша)
Университет «Туран»

7–11 октября 2026 г.

Алматы, 2026

УДК 004.4:519.6 (063)
ББК 32.973.202
И 74

Главный редактор:
академик НАН РК, доктор физико-математических наук, профессор
Калимолдаев М.Н.

Ответственные редакторы:
декан механико-математического факультета КазНУ имени аль-Фараби, к.ф.-м.н., профессор **Мансурова М.Е.**
ученый секретарь ИИВТ КН МНВО РК, PhD, ассоц. профессор **Усатова О.А.**
старший научный сотрудник ИИВТ КН МНВО РК, PhD, ассоц. профессор **Зиятбекова Г.З.**
старший научный сотрудник ИИВТ КН МНВО РК, PhD **Аршидинова М.Т.**
заведующая кафедрой «Искусственный интеллект и Big Data» КазНУ имени аль-Фараби, PhD **Сарсембаева Т.С.**
секретарь конференции, КазНУ имени аль-Фараби **Айдарова Л.Н.**

И 74 Информатика и прикладная математика: Материалы XI Междунар. науч.-практ. конф. (07.10 – 11.10.2026 г.). В 2-х т. Т. 2. – Алматы, 2026. – 358 с.

ISBN 978-601-228-616-8 (т. 2)
ISBN 978-601-228-614-4 (общ.)

В сборнике опубликованы доклады, представленные учеными Республики Казахстан, Российской Федерации, Республики Узбекистан, Азербайджанской Республики, Кыргызской Республики, Турецкой Республики, Республики Польша, Венгрии, Индии, Греции, Китайской Народной Республики и других стран.

Рассмотрены актуальные вопросы математического моделирования, оптимизации и вычислительного анализа, искусственного интеллекта, больших данных и интеллектуальных систем, инноваций в области управления компьютерными науками и инженерии, кибербезопасности, конфиденциальности и цифрового доверия.

Материалы сборника предназначены для научных работников, докторантов, магистрантов, а также студентов старших курсов.

УДК 004.4:519.6 (063)
ББК 32.973.202

ISBN 978-601-228-616-8 (т. 2)
ISBN 978-601-228-614-4 (общ.)

© Институт информационных и вычислительных технологий КН МНВО РК, 2026
© Казахский национальный университет имени аль-Фараби, 2026

КРАТКАЯ ПРОГРАММА
XI Международной научно-практической конференции
«Информатика и прикладная математика»
«Computer Science and Applied Mathematics»

Дата проведения: **7–11 октября 2026 г.**

Время: **09:00–18:00**

Место проведения: **Библиотека КазНУ имени аль-Фараби,**
зал Абая (ауд. № 308), ауд. № 408
г. Алматы, проспект аль-Фараби, 71/27

Среда, 7 октября	
09:00–18:00	Встреча и прибытие зарубежных участников
Четверг, 8 октября	
09:30–10:00	Официальная регистрация участников конференции и встреча гостей
10:00–11:00	**Приветственное слово**
11:00–13:10	**Пленарное заседание**
13:10–14:30	Перерыв на обед
14:30–16:00	Секционные заседания
16:00–18:00	Круглый стол «Искусственный интеллект и языковые технологии: от научных основ к технологиям будущего»
Пятница, 9 октября	
10:00–12:10	**Пленарное заседание**
12:10–14:30	Перерыв на обед
14:30–17:10	Секционные заседания
Суббота, 10 октября	
09:00–18:00	День культурной и социальной программы
Воскресенье, 11 октября	
Завершение работы конференции, отъезд участников	

ОРГАНИЗАТОРЫ И КОМИТЕТЫ КОНФЕРЕНЦИИ
XI Международной научно-практической конференции
«Информатика и прикладная математика»
«Computer Science and Applied Mathematics»

Организаторы:

- Казахский национальный университет имени аль-Фараби (г. Алматы, Республика Казахстан)
- Институт информационных и вычислительных технологий КН МНВО РК (г. Алматы, Республика Казахстан)
- Люблинский технический университет (г. Люблин, Польша)
- Университет «Туран» (г. Алматы, Республика Казахстан)

Программный комитет

Председатель Международного программного комитета:

- Калимолдаев Максат Нурадилович, академик НАН РК, д.ф.-м.н., профессор, Республика Казахстан

Сопредседатели Международного программного комитета:

- Бекманова Гульмира Тылеубердиевна, к.т.н., доктор PhD, доцент, Республика Казахстан
- Ибраимов Маргулан Касенович, PhD, доцент, Республика Казахстан
- Разакова Дина Ибрагимовна, PhD, доцент, Республика Казахстан

Заместители Председателя Международного программного комитета:

- Мансурова Мадина Есимхановна, к.ф.-м.н., профессор, Республика Казахстан
- Усатова Ольга Александровна, PhD, ассоц. профессор, Республика Казахстан
- Зиятбекова Гулзат Зиятбековна, PhD, ассоц. профессор, Республика Казахстан

Члены Международного программного комитета:

- Abdul Razaque, PhD, USA
- Abdullayev Fahreddin, professor, Turkey
- Aijaz Shaikh, professor, Finland
- Andrzej Smolarz, professor, Poland
- Dietmar Beyer, PhD, Germany
- Dmitrishin Roman, professor, Ukraine
- Hazar Elman, professor, Turkey

- Karimov Adalat, professor, Azerbaijan
- Keylan Alimhan, professor, Japan
- Maksim Iavich, PhD, professor, Georgia
- Michele Pagano, professor, Italy
- Mohamed Othman, professor, Malaysia
- Mosavi Amir, professor, Hungary
- Müslüm Arici, professor, Turkey
- Ravil Muhamediev, professor, Latvia
- Айда-заде Камиль Раджабович, д.ф.-м.н., профессор, член-корреспондент НАН Азербайджана
- Мальцева Светлана Валентиновна, д.т.н., профессор, Россия
- Марченко Михаил Александрович, д.т.н., профессор, Россия
- Матерухин Андрей Викторович, д.т.н., профессор, Россия
- Медведев Сергей Борисович, д.ф.-м.н., профессор, Россия
- Посыпкин Михаил Анатольевич, член-корреспондент РАН, Россия
- Родионов Алексей Сергеевич, д.т.н., профессор, Россия
- Тузиков Александр Васильевич, д.т.н., профессор, член-корреспондент НАН Беларуси
- Алгазы Кунболат Тилеуханулы, PhD, Республика Казахстан
- Аманжолова Сауле Токсановна, к.т.н., ассоц. профессор, Республика Казахстан
- Амиргалиева Салтанат Нурадиловна, д.ф.-м.н., профессор, Республика Казахстан
- Амирханова Гульшат Аманжоловна, доктор PhD, Республика Казахстан
- Ахметжанов Максат Аканович, PhD, Республика Казахстан
- Ахметов Искандер Рафаилович, PhD, Республика Казахстан
- Бабенко Татьяна Васильевна, д.т.н., профессор, Республика Казахстан
- Боймуродов Фаррух Фарход угли, PhD, Республика Узбекистан
- Байтенова Лаура Маратовна, к.э.н., профессор, Республика Казахстан
- Барахнин Владимир Борисович, д.т.н., профессор, Россия
- Бегимбаева Енлик Ериковна, PhD, Республика Казахстан
- Бектемесов Мактагали Абдимажитович, д.ф.-м.н., профессор, Республика Казахстан
- Бектемысова Гульнара Умиткуловна, к.т.н., ассоц. профессор, Республика Казахстан
- Бельгибаев Бауржан Абдрахимович, д.т.н., доцент, Республика Казахстан

- Галбаев Жалалидин Токтобаевич, д.т.н., профессор, Кыргызская Республика
- Дарибаев Беимбет Серикович, PhD, ассоц. профессор, Республика Казахстан
- Жартыбаева Макпал Галымбековна, PhD, Республика Казахстан
- Жукабаева Тамара Кукеновна, PhD, Республика Казахстан
- Иманкулов Тимур Сакенович, PhD, ассоц. профессор, Республика Казахстан
- Искаков Казизат Такуадинович, д.ф.-м.н., профессор, Республика Казахстан
- Исмаилов Бактыбек Искакович, д.т.н., профессор, Кыргызская Республика
- Ишанходжаев Гайрат, д.т.н., профессор, Республика Узбекистан
- Калинин Алексей Анатольевич, PhD, доцент, Республика Казахстан
- Калижанова Алия Уалиевна, к.ф.-м.н., профессор, Республика Казахстан
- Капалова Нурсулу Алдажаровна, к.т.н., ассоц. профессор, Республика Казахстан
- Карюкин Владислав Игоревич, PhD, Республика Казахстан
- Киселева Ольга Владимировна, к.т.н., ассоц. профессор, Республика Казахстан
- Козбакова Айнур Холдасовна, PhD, ассоц. профессор, Республика Казахстан
- Кошеков Кайрат Темирбаевич, д.т.н., профессор, академик НАН РК, Республика Казахстан
- Касымбек Нурислам, PhD, Республика Казахстан
- Лисневский Ростислав Валерьевич, PhD, ассоц. профессор, Республика Казахстан
- Мазаков Талгат Жакупович, д.ф.-м.н., профессор, Республика Казахстан
- Мусабаев Рустам Рафикович, к.т.н., ассоц. профессор, Республика Казахстан
- Мусиралиева Шынар Женисбековна, к.ф.-м.н., доцент, Республика Казахстан
- Найзабаева Ляззат Кадыралиевна, д.т.н., профессор, Республика Казахстан
- Назирова Эльмира Шодмоновна, д.т.н., профессор, Республика Узбекистан
- Нуралиев Фахриддин Муродиллаевич, д.т.н., профессор, Республика Узбекистан
- Омаров Батырхан, PhD, Республика Казахстан
- Пак Александр Александрович, к.т.н., профессор, Республика Казахстан
- Пак Иван Тимофеевич, д.т.н., профессор, Республика Казахстан
- Рахимова Диана Рамазановна, PhD, Республика Казахстан

- Равшанов Нормахмад, д.т.н., профессор, Республика Узбекистан
- Сакан Кайрат, PhD, Республика Казахстан
- Сатыбалдина Дина Жагипаровна, к.ф.-м.н., ассоц. профессор, Республика Казахстан
- Сеилова Нургуль Абадуллаевна, к.т.н., ассоц. профессор, Республика Казахстан
- Сеилов Шахмаран Журсинбекович, д.э.н., профессор, Республика Казахстан
- Сейткулов Ержан Нураханович, к.ф.-м.н., ассоц. профессор, Республика Казахстан
- Сембиев Ордабай Зайтаевич, д.т.н., профессор, Республика Казахстан
- Тастанова Саида Алдияровна, д.т.н., PhD, доцент, Республика Узбекистан
- Ташатов Нурлан Наркенович, к.ф.-м.н., PhD, ассоц. профессор, Республика Казахстан
- Тиреуов Канат Маратович, академик НАН РК, д.э.н., Республика Казахстан
- Торобеков Бекжан Торобекович, д.т.н., профессор, Кыргызская Республика
- Тукеев Уалшер Ануарбекович, д.т.н., профессор, Республика Казахстан
- Турарбек Асем, PhD, Республика Казахстан
- Тусупова Сауле Амангельдиевна, д.т.н., профессор, Республика Казахстан
- Утегенова Анара Урантаевна, PhD, ассоц. профессор, Республика Казахстан
- Хомпыш Ардабек, PhD, Республика Казахстан
- Шормакова Асем Ноябревна, PhD, Республика Казахстан
- Эшкараева Нархол Гузаровна, к.т.н., профессор, Республика Узбекистан

Ученые секретари конференции:

- Аршидинова Мукаддас Тургановна, PhD, Республика Казахстан
- Сарсембаева Талшын Сагдатбековна, PhD, Республика Казахстан

Организационный комитет

Председатели:

- Усатова Ольга Александровна, PhD, ассоц. профессор, Республика Казахстан
- Мансурова Мадина Есимхановна, к.ф.-м.н., профессор, Республика Казахстан

Члены Организационного комитета:

- Айдарова Л.Н., Республика Казахстан
- Абдуали Б.А., PhD, Республика Казахстан
- Аспантаев А.Б., Республика Казахстан
- Варенников А.В., Республика Казахстан
- Лизунов О.А., Республика Казахстан
- Меркебаев А.Г., Республика Казахстан
- Мухитова А.А., Республика Казахстан
- Ткачев К.В., Россия
- Толеу А., Республика Казахстан
- Шахмаев Р.А., Республика Казахстан

Соорганизаторы конференции

Дорогие друзья, уважаемые коллеги, организаторы и участники XI Международной научно-практической конференции «Информатика и прикладная математика»!

От имени Международного программного комитета и организаторов сердечно приветствую всех участников **XI Международной научно-практической конференции «Информатика и прикладная математика»**! Наша конференция, которая в этом году проводится уже в одиннадцатый раз, стала признанной научной площадкой для обмена идеями и результатами исследований в области математического моделирования, искусственного интеллекта, анализа данных, информационной безопасности и цифровых технологий. Выражаю искреннюю благодарность Казахскому национальному университету имени аль-Фараби, Институту информационных и вычислительных технологий КН МНВО РК, Люблинскому техническому университету и Университету «Туран» за совместную работу по подготовке и проведению конференции.

Мы живем в период, когда информационные технологии и прикладная математика определяют темпы развития экономики, науки и общества. Искусственный интеллект переходит из лабораторий в повседневную практику – в медицину, сельское хозяйство, промышленность, энергетику, образование и государственное управление. Вместе с новыми возможностями возникают и новые вызовы: надежность и объяснимость интеллектуальных систем, защита персональных данных, кибербезопасность критической инфраструктуры, сохранение языкового и культурного многообразия в эпоху больших языковых моделей. Ответы на эти вопросы невозможны без фундаментальной математической подготовки, междисциплинарного сотрудничества и открытого научного диалога.

Развитие искусственного интеллекта и цифровизация экономики определены в числе стратегических приоритетов Республики Казахстан. Наращивание потенциала отечественной науки, подготовка высококвалифицированных кадров и ускоренное внедрение результатов исследований в практику – важнейшие задачи, решению которых призвана служить и наша конференция.

В этом году в работе конференции принимают участие ученые и специалисты из Казахстана, России, Узбекистана, Азербайджана, Кыргызстана, Турции, Польши, Венгрии, Индии, Греции, Китая и других стран. Программа включает пленарные доклады ведущих ученых, работу четырех секций – «Математическое моделирование, оптимизация и вычислительный анализ», «Искусственный интеллект, Big Data, наука о данных и интеллектуальные системы», «Инновации в области управления компьютерными науками и инженерии», «Кибербезопасность, конфиденциальность и цифровое доверие», а

также круглый стол «Искусственный интеллект и языковые технологии: от научных основ к технологиям будущего». Особенно радует, что значительную часть докладов представляют молодые ученые, докторанты и магистранты – именно им предстоит определять будущее казахстанской и мировой науки.

Проведение наших ежегодных встреч стало доброй традицией. С каждым годом растет число участников, повышается качество выступлений и публикаций, расширяется география научного сотрудничества. Уверен, что и нынешняя конференция станет местом плодотворных дискуссий, новых научных контактов и совместных проектов, а идеи, высказанные здесь, найдут воплощение в исследованиях и практических разработках.

Уважаемые участники и организаторы конференции! Позвольте выразить Вам глубокую благодарность за участие, поддержку и развитие нашей конференции. Желаю всем плодотворной работы, интересных докладов, творческих успехов и новых научных достижений!

Председатель Международного
программного комитета,
академик НАН РК, д.ф.-м.н., профессор **Калимолдаев М.Н.**

СОДЕРЖАНИЕ

стр.

Секция 3. Инновации в области управления компьютерными науками и инженерии

1. **Калижанова А.У., Утегенова А.У., Кунелбаев М.М., Дәруіш С.З., Иманбекова У.Н.** 16
Моделирование движения отделяющейся ступени ракеты-носителя в MATLAB/Simulink

2. **Ziyatbekova G.Z., Akhmetova A.M., Zordunov T.K., Sydykova M.M., Sagyntay O.A.** 23
PSO-optimized PID control in gas condensate absorption

3. **Duran V., Hazar E.** 28
Entropy-guided adaptive scaffolding in mathematics education: a Bayesian decision model

4. **Duran V., Hazar E.** 37
The VERIFY cycle: a verification-first pedagogy for generative AI in mathematical modelling

5. **Nurgazy T., Amirkhanova G., Amirkhanov B., Rustemov A., Liu H.** 43
Integrating web-based augmented reality into industrial dashboards: a case study on real-time factory equipment monitoring

6. **Zhussip A., Ziyatbekova G., Sarsembayeva T., Akhmetova A., Tengaeva A., Tasbolatuly N.** 50
Forecasting crop yields using XGBoost and Bayesian hyperparameter optimization

7. **Мадьярова Г.А., Орынбаева А.С.** 57
Дарынды оқушылармен жұмыс істеуде инновациялық педагогикалық технологияларды қолдану

8. **Максутова Ш.У.** 64
Жасанды интеллект негізінде сүт безі ауруларын ерте анықтаудың интеллектуалды модельдерін әзірлеу

9. **Зиятбекова Г.З., Абдуали Б.А., Жаныбекулы С., Дудин А.В., Андронова Т.А.** 73
Разработка интеллектуальной системы цветовой приоритизации и планирования

10. **Самбетбаева А.К., Қажыбек А.М., Байкувеков М.Б.** 78
Ghost-CA-YOLO архитектурасы негізінде жол зақымдануларын нақты уақыт режимінде анықтау және бағалау

11. **Ақылаев Ж.А.** 84
Қаржылық-технологиялық экожүйелердегі интеллектуалды агенттік шешім қабылдау жүйелері

12. **Мутанов Г.М., Усатова О.А., Науменко В.В., Карюкин В.И., Төреқұл С.М., Аманжолова С.Т.** 90
Гибридная интеллектуальная модель формальной проверки заявок на научные и научно-технические проекты

13. **Торекул С.М., Мутанов Г.М.** 96
Пилотная апробация модели классификации этапов НИОКР на основе формализованных правил

14. **Нерезов В.К.** 104
Разработка облачной платформы для автоматической генерации веб-сайтов и презентаций на основе больших языковых моделей

15. **Раимбаев М.С.** 110
Облачные технологии как основа современной IT-инфраструктуры

16. **Сатыбалдин А.Т.** 117
Исследование и разработка онлайн-платформы для проведения тестирований и опросов в режиме реального времени с использованием машинного обучения

17. **Исабаева Д.Н., Акупова Д.Н.** 122
Coursera онлайн-курстарын дербестендіріп таңдауға арналған интеллектуалды ұсыным жүйесін әзірлеу

18. **Попов А.Г.** 128
Анализ современных методов прогнозирования технического обслуживания оборудования на основе логов для системы LogOracle

19. **Amirkhanova G., Amirkhanov B., Raeva A., Koksunbaev A., Kenzhekhanuly T., Liu H.** 133
Remote electricity metering nodes with Wi-Fi and LoRaWAN channels for the OpenEgiz digital twin platform

20. **Нурхожаев Ж.М., Амирханова Г.А.** 140
Интеллектуальный мониторинг и управление микроклиматом теплиц с применением искусственного интеллекта и промышленного интернета вещей: систематический обзор

21. **Калимолдаев М.Н., Толегенова Г.Б.** 157
Деректерді диагностикалаудан интеллектуалды платформаға дейін: жаңартылатын энергетикадағы біртұтас зерттеу желісі

22. **Ishniyazov O.O., Babajanov M.R., Shokirov Sh.Sh.** 162
Automatic linkage of MARC bibliographic records using adaptive blocking and hybrid similarity features

23. **Оган Н., Исабаева Д.Н.** 168
технологияларын қолдану

24. **Рсымбетов К., Есмаханова К.** 169
Разработка и экспериментальное внедрение адаптивной образовательной системы на основе искусственного интеллекта для подготовки IT-специалистов

25. **Джунусалиев Н.Дж., Тойбаева Ж.Дж., Жамгырчиева Г.Т.** 174
Разработка кыргызско-русского электронного словаря технических терминов «Техносөздүк» и перспективы его развития с применением технологий

26. **Kemelbekov N., Rakhimova D.R.** 181
Machine learning-driven adaptive traffic management for multi-cluster Kubernetes service mesh via Gateway API

27. **Кабдушев Ш.Б., Сергеев А.С., Нуралиев Ф.М., Байпакбаева С.Т., Мун Г.А.** 184
Искусственный интеллект в организации научного поиска: подходы к исследованию нейроморфных свойств и вычислительного потенциала полимеров

28. **Жамбылхан С.Е.** 193
Контейнеризация и оркестрация как основа cloud-native разработки программного обеспечения

Секция 4. Кибербезопасность, конфиденциальность и цифровое доверие

29. **Алғазы К.Т., Әлімжан Е.Ж., Сақан Қ.С.** 201
FLUX-128 жеңілсалмақты криптографиялық хеш-функциясының лавиндық сипаттамаларын зерттеу

30. **Хаумен А., Алгазы К.Т.** 213
Исследование влияния параметров циклических сдвигов на диффузионные свойства блочного шифра FEATHER-RX

31. **Сакан К.С., Нысанбаева С.Е., Маро Е.А., Накипова Э.Б.** 220
Оценка эффективности и сравнительный анализ цифровой подписи на основе дерева Verkle

32. **Капалова Н.А., Лизунов О.А., Дюсенбаев Д.С., Остапенко В.В., Алдажар Ж.К.** 231
Разработка и исследование алгоритма шифрования SAUT для защищенного обмена между IoT-устройствами на примере STM32F769I-DISCO

33. **Сайманов И.М., Сапарниязов Б.А.** 238
Защита данных IoT (интернета вещей) на основе алгоритмов гомоморфного шифрования

34. **Аширбаев Н.Қ., Карюкин В.И.** 248
Вероятностный анализ многофакторной аутентификации для защиты данных пациентов в клинических информационных системах

35. **Dendukuri S.V.** 255
Federated learning for privacy-preserving ESG risk assessment in cloud-native

36. **Степаненко Ж.О., Муканова Ж.А.** 266
Финансовая киберпреступность в Республике Казахстан: угрозы и методы защиты

37. **Турарбек Ә.Т., Темиргазиева Ш.** 272
Аудио дипфейктерді анықтау үшін деректер жиынтықтарын зерттеу және нейрондық желі негізінде жіктеу

38. **Ташатов Н.Н., Жүзбаев С.С., Сарсенбай М.С.** 282
Обеспечение информационной безопасности информационно-аналитической системы прогнозирования свойств композитных материалов на основе многокритериального принятия решений

39. **Жақсылық Д.О., Кенжебаева М.О.** 288
смишинг шабуылдарын анықтау және талдау жүйесін құру

40. **Al Jabid, Shihab A.H., Kenzhebayeva M.O.** 294
Measuring jailbreak robustness of small language models with a reproducible authorization-gated evaluation framework

41. **Khamidulla A.** 301
Modeling and detection of known and zero-day network attacks using AI/ML algorithms

42. **Kabilov R.I., Reypnazarov E.N., Saymanov I.M.** 318
Detecting and mitigating load telemetry falsification in GRU-based predictive resource allocation

43. **Дюсенбаев Д.С., Хомпыш А., Ищукова Е.А.** 326
Құпияны бөлісу схемалары: әдістері мен қолданылуы

44. **Ким А.В., Усатова О.А.** 330
Гибридная модель обнаружения сетевых вторжений на основе ансамбля классификаторов и CNN-BiLSTM-Attention

45. **Мекебаев Н.О., Ашим Т.Т.** 338
Мемлекеттік ақпараттық ресурстар қауіпсіздігін бағалауда аудит жүйелерінің тиімділігін талдау

46. **Мустафина К.С.** 347
Защита персональных данных в современных информационных системах

47. **Капалова Н.А., Ергеш Н.С.** 353
Бет-әлпет биометриялық шаблондарын таралған сақтауға арналған құпияны табалдырықты бөлу схемаларын салыстырмалы талдау

УДК 004.031

МОДЕЛИРОВАНИЕ ДВИЖЕНИЯ ОТДЕЛЯЮЩЕЙСЯ СТУПЕНИ РАКЕТЫ-НОСИТЕЛЯ В MATLAB/SIMULINK

А. У. Калижанова[1,2], А.У. Утегенова[1,3], М.М. Кунелбаев[1,4], С.З. Дәруіш[1,4],
У.Н. Иманбекова [1,2]

[1]*Институт информационных и вычислительных технологий КН МНВО РК, Казахстан*

[4]*Алматинский университет энергетики и связи им.Г. Даукеева, Казахстан*

[3]*Satbayev University, Казахстан*

[3]*Казахский национальный университет им. Аль-Фараби, Казахстан*

e-mail: aliya_kalizhanova@mail.ru

Аннотация. *Разработана компактная модульная модель движения отделяющейся ступени ракеты-носителя в MATLAB/Simulink. Модель объединяет стандартную атмосферу, изменение гравитации с высотой, пространственный ветер, аэродинамические силы и моменты, шестистепенные уравнения движения и интегрирование переменных состояния. Расчётный эксперимент выполнен для участка длительностью 30 с при снижении с 10 000 до 6370 м. Получены изменения координат, линейных и угловых скоростей, аэродинамического сопротивления и бокового отклонения. Максимальная скорость составила 300 м/с, результирующая скорость ветра — 7,19 м/с, а поперечное смещение достигло 207 м. Предложенная архитектура предназначена для оперативного прогнозирования траектории и вероятной зоны падения при ограниченных вычислительных ресурсах.*

Ключевые слова: *ракета-носитель, отделяющаяся ступень, 6-DOF, MATLAB/Simulink, траектория, зона падения.*

1 Введение

Современное моделирование движения ракет основано на совместном описании поступательной и вращательной динамики, атмосферных параметров, гравитационного поля, аэродинамических воздействий, ветровых возмущений и изменения состояния объекта. Особую задачу представляет движение отделяющейся ступени после завершения управляемого участка. На этом этапе траектория формируется преимущественно под действием силы тяжести, сопротивления воздуха и ветра, а неопределённость параметров быстро увеличивает ошибку определения точки падения. В исследовании [1] разработана шестистепенная модель движения неуправляемой исследовательской ракеты в MATLAB/Simulink. Авторы рассмотрели полёт при отсутствии ветра и при воздействии ветра различных направлений, а также сопоставили полученные результаты с расчётами в OpenRocket. Использование аэродинамических данных Missile DATCOM позволило учитывать изменение угла атаки и динамическую устойчивость ракеты. Авторами исследования [2] представлена модель запуска и управления двухступенчатой ракетой, основанная на шестистепенных уравнениях движения. Реализация в MATLAB/Simulink позволила объединить динамику отдельных ступеней, управляющие воздействия и последовательность этапов полёта в единой вычислительной среде. В публикации [3] предложен программный инструмент для моделирования шестистепенной динамики ракеты-носителя. Разработанная модель охватывает движение от момента запуска до орбитального выведения либо возвращения и включает отдельные модули окружающей среды, аэродинамики, силовых воздействий и движения центра масс. Исследование [4] посвящено разработке высокоточной динамической модели исследовательской ракеты, предназначенной для

предварительного проектирования. В модели учитываются аэродинамические коэффициенты, изменение массы, характеристики атмосферы, ветровые порывы и турбулентность, включая модели Dryden и von Kármán. В статье [5] сформирована нелинейная 6-DOF-модель ракеты для аппаратно-программных испытаний системы управления по схеме Hardware-in-the-Loop. Модель обеспечивает вычисление пространственного положения, линейной скорости, ориентации и угловых скоростей ракеты в реальном времени. В рамках исследования [6] выполнено сравнение двух подходов к аэродинамическому моделированию с использованием шестистепенной модели движения, реализованной в MATLAB/Simulink. Установлено, что метод определения аэродинамических коэффициентов существенно влияет на расчёт сил, моментов и пространственной траектории объекта. Авторы публикации [7] предложили комплексную модель и метод оптимизации траектории твердотопливной ракеты. В MATLAB/Simulink были реализованы шестистепенные уравнения движения, аэродинамические зависимости и оптимизационные процедуры, позволяющие исследовать влияние конструктивных параметров на траекторию полёта. В исследовании [8] рассмотрено физическое моделирование многоразовых ракет с использованием объектно-ориентированного подхода. Разработанная библиотека содержит модели корпуса, двигательной установки, исполнительных механизмов и внешней среды и предназначена для проверки алгоритмов навигации, стабилизации и управления. Публикация [9] посвящена управляемой посадке первой ступени ракеты-носителя с применением глубокого обучения с подкреплением. Обучение и тестирование алгоритма выполнены на нелинейном шестистепенном симуляторе, учитывающем поступательную и вращательную динамику первой ступени. В исследовании [10] разработана система определения и управления ориентацией исследовательской ракеты с отклоняемым вектором тяги. Для проектирования регулятора была сформирована нелинейная 6-DOF-модель динамики и кинематики ракеты.

Целью исследования является разработка модульной модели движения отделяющейся ступени ракеты-носителя в среде MATLAB/Simulink для расчёта её пространственной траектории, параметров движения и прогнозируемой зоны падения с учётом гравитации, атмосферных условий, аэродинамического сопротивления и ветровых возмущений.

2 Материалы и методы

Функциональная архитектура включает начальные условия, стандартную атмосферу, модель гравитации, пространственный ветер, аэродинамический расчёт, преобразование сил и моментов, 6-DOF-динамику и интегратор состояния. Входными данными служат начальные координаты, скорость, углы ориентации, угловые скорости, масса, площадь характерного сечения и аэродинамические коэффициенты.

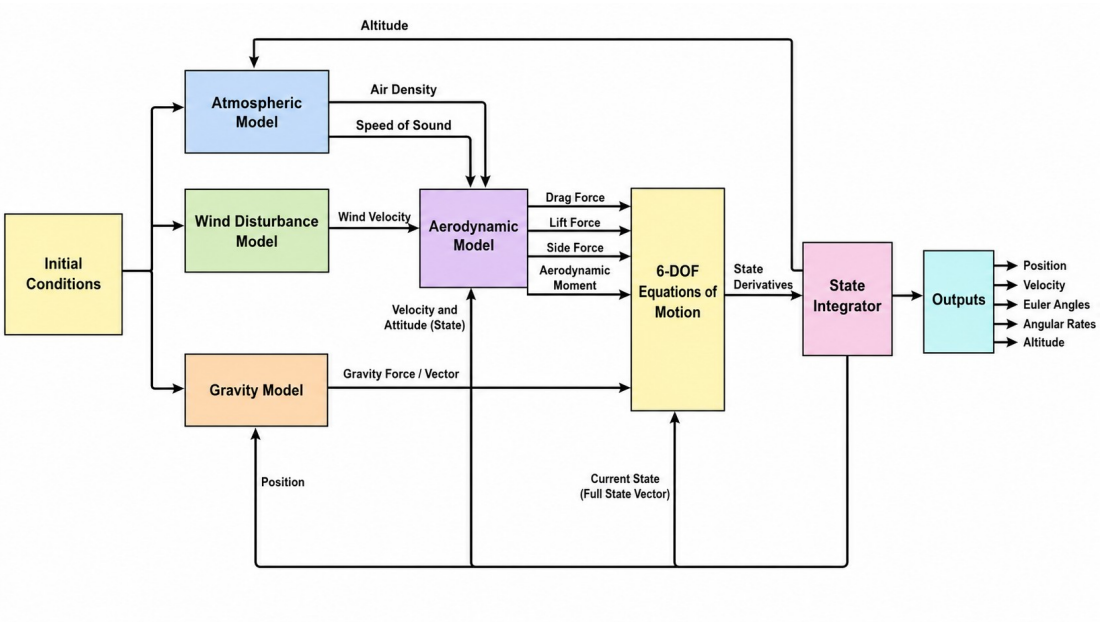


Рис. 1 Интегрированная архитектура моделирования движения отделяющейся ступени

Как видно из представленного рисунка 1, текущие координаты, скорости, углы ориентации и угловые скорости используются в контурах обратной связи для обновления атмосферных, гравитационных и аэродинамических параметров. Аэродинамические силы и моменты совместно с гравитационным воздействием поступают в блок шестистепенных уравнений движения, где формируются производные переменных состояния. После интегрирования определяются положение, скорость, углы Эйлера, угловые скорости и высота отделяющейся ступени. Такая модульная структура позволяет независимо настраивать и проверять отдельные компоненты модели, а также расширять архитектуру дополнительными блоками управления, неопределённостей и прогнозирования зоны падения.

Реализация модели в MATLAB/Simulink

Атмосферная подсистема вычисляет температуру, давление, плотность и скорость звука в зависимости от геометрической высоты. Гравитационная подсистема формирует силу тяжести с учётом расстояния до центра Земли. Модель Драйдена задаёт продольную, поперечную и вертикальную составляющие турбулентного ветра. Аэродинамический блок определяет число Маха, коэффициенты сил и моментов, динамическое давление и результирующие нагрузки.

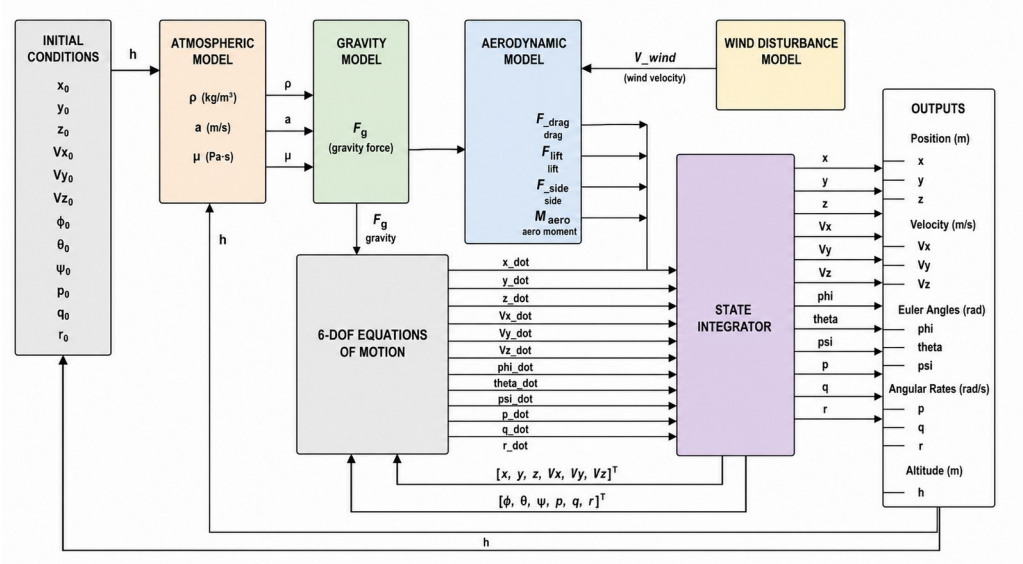


Рис. 2 Полная MATLAB/Simulink-архитектура модели

На рисунке 2 представлена полная MATLAB/Simulink-архитектура моделирования движения отделяющейся ступени. Подсистема атмосферной модели вычисляет плотность воздуха и скорость звука, гравитационная модель формирует вектор силы тяжести, а модель ветровых возмущений задаёт компоненты скорости ветра, поступающие в аэродинамический блок

Расчётный интервал составляет 30 с. Начальная высота принята равной 10 000 м, начальная полная скорость — 121,7 м/с, масса ступени — 1000 кг. Компоненты ветра изменяются во времени, а максимальная результирующая скорость ветра достигает 7,19 м/с. Тяга после отделения не учитывается; масса и тензор инерции на рассматриваемом коротком интервале считаются постоянными.

Модель регистрирует высоту, пройденное расстояние, три координаты, три компоненты линейной скорости, углы ориентации, угловые скорости, число Маха, динамическое давление, силу сопротивления, силу тяжести и боковое отклонение. Такой набор параметров достаточен для восстановления пространственной траектории и последующего построения прогнозируемой зоны падения.

Модульная организация позволяет независимо заменять атмосферные профили, ветровые сценарии и таблицы аэродинамических коэффициентов. При наличии телеметрии блок начальных условий может обновляться по измеренным координатам и скоростям, что создаёт основу для применения расширенного фильтра Калмана или фильтра частиц без перестройки физического ядра модели.

Результаты

За 30 с расчёта высота ступени уменьшилась с 10 000 до 6370 м, а продольная координата увеличилась до 4860 м. Поперечное смещение достигло 207 м. Скорость сначала возрастала под действием гравитации, достигла 300 м/с на 25-й секунде, а затем снизилась до 280 м/с вследствие роста аэродинамического сопротивления.

Таблица 1. Основные результаты моделирования

Показатель	**Начальное значение**	**Максимальное/конечное значение**	**Единица измерения**
Высота	10 000	6370	м
Полная скорость	121,7	300 / 280	м/с
Продольная координата x	0	4860	м
Поперечное смещение y	0	207	м
Результирующая скорость ветра	4,0	7,19	м/с
Сила аэродинамического сопротивления	—	3950	Н

В таблице 1 представлены основные результаты моделирования движения отделяющейся ступени ракеты. За расчётный период высота уменьшилась с 10 000 до 6370 м, а полная скорость достигла максимального значения 300 м/с и к концу расчёта снизилась до 280 м/с. Продольная координата составила 4860 м, поперечное смещение — 207 м, результирующая скорость ветра — 7,19 м/с, а максимальная сила аэродинамического сопротивления достигла 3950 Н.

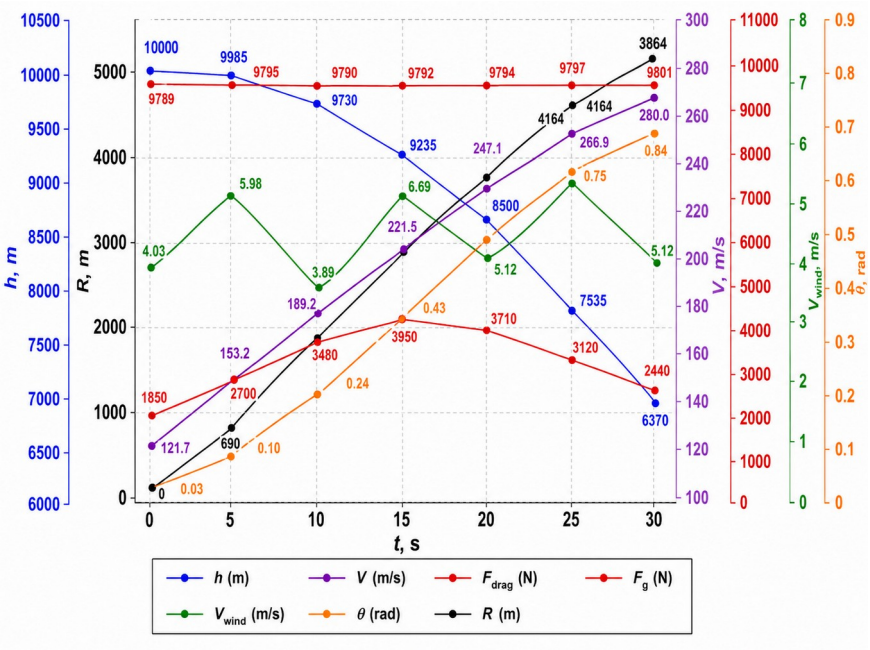


Рис. 3 Совместное изменение основных параметров движения ступени

На рисунке 3 представлено совместное изменение высоты, скорости, аэродинамической силы сопротивления, силы тяжести, скорости ветра, угла тангажа и пройденного расстояния в течение 30 с движения ступени. Высота уменьшается с 10 000 до 6370 м, тогда как скорость возрастает со 121,7 до 280,0 м/с, а пройденное расстояние достигает 4864 м. Сила сопротивления увеличивается до максимального значения 3950 Н на 15-й секунде, после чего снижается до 2440 Н, в то время как сила тяжести изменяется незначительно и остаётся в диапазоне 9789–9801 Н. Угол тангажа постепенно возрастает до 0,84 рад, а скорость ветра изменяется неравномерно, достигая максимума 7,19 м/с на 25-й секунде.

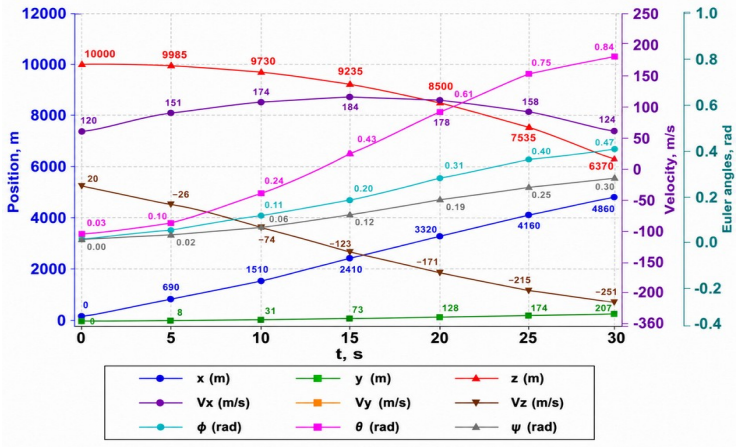


Рис. 4 Изменение координат, скоростей и углов ориентации в течение 30 с

На рисунке 4 показано изменение координат, скоростей и углов ориентации ступени в течение 30 с движения. Продольная координата xвозрастает от 0 до 4860 м, поперечное смещение y— до 207 м, тогда как вертикальная координата zуменьшается с 10 000 до 6370 м. Скорость V_xсначала увеличивается до 184 м/с, а затем снижается до 124 м/с, в то время как V_zизменяется от 20 до -251м/с, что свидетельствует о переходе ступени к интенсивному снижению. Углы φ, θи ψпостепенно возрастают, подтверждая существенное изменение пространственной ориентации аппарата.

Заключение

Разработана модульная MATLAB/Simulink-модель движения отделяющейся ступени, объединяющая атмосферную, гравитационную, ветровую, аэродинамическую и шестистепенную динамические подсистемы. В расчётном сценарии за 30 с высота снизилась с 10 000 до 6370 м, продольная координата достигла 4860 м, поперечное смещение — 207 м, а скорость — максимального значения 300 м/с. Результирующая скорость ветра составила до 7,19 м/с, сила сопротивления — до 3950 Н.Полученные результаты подтверждают работоспособность замкнутой архитектуры и её пригодность для предварительного прогнозирования траекторий. В дальнейшем модель целесообразно дополнить обработкой телеметрии, вероятностной оценкой неопределённости и экспериментальной верификацией. Дальнейшая работа предусматривает подключение расширенного фильтра Калмана и фильтра частиц для коррекции состояния по телеметрии, вероятностное моделирование ветра и начальных ошибок методом Монте-Карло, учёт вращения Земли и неоднородности атмосферы, а также экспорт траекторий в геоинформационную систему. Это позволит перейти от одной расчётной линии к вероятностной карте зоны падения с доверительными границами. Исследования проведены при поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан в рамках гранта № AP23488291 «Разработка многофункционального ресурса экологической паспортизации районов падения отделяющихся частей ракет-носителей методом адаптивного представления интерактивных ГИС».

Литература

1. W. L. de Oliveira Junior, H. A. Fazzolari, C. M. Freire, and C. A. de Paiva Carvalho. 6 Degrees of Freedom Simulation of an Unguided Sounding Rocket Using Matlab/Simulink. SAE Technical Paper 2023-36-0095, SAE Brasil 2023 Congress. DOI: 10.4271/2023-36-0095.

2. K. Devaki, A. Hussain, S. Reza. Six Degree of Freedom Simulation of Control Systems and Launch of a Two-Staged Rocket. Proceedings of the 8th International Conference on Signal Processing and Integrated Networks. 2021, pp.856-861. IEEE. DOI: 10.1109/SPIN52536.2021.9566037

3. G. Silveira, V. Carrara. A Six Degrees-of-Freedom Flight Dynamics Simulation Tool of Launch Vehicles. Journal of Aerospace Technology and Management. 2015. vol. 7, №2, pp. 231-239. DOI: 10.5028/jatm.v7i2.433.

4. H. Li, X.-G. Jing, and L.-M. Xu. The Simulation of 6-DOF Motion of Launch Vehicle Based on Matlab/Simulink. Journal of Astronautics**.** vol. 26, no. 5, pp. 616–619, 2005**.**

5. M. Afshar, M. F. Çakır, M. Bayraktar. Trajectory Modeling for High-Power Rockets Using MATLAB/Simulink. Journal of Engineering Sciences and Design, vol. 13, no. 3, pp. 817–828, 2025, doi: 10.21923/jesd.1653290.

6. M. Tayefi, A. Mohammadi, M. Ebrahimi. Development of a Space Launch Vehicle 6DoF Multipurpose Simulation Software. 18th *AIAA/3AF* International Space Planes and Hypersonic Systems and Technologies Conference. Paper AIAA 2012-5883. 2012. https://doi.org/10.2514/6.2012-5883.

7. Zhang Wenpeng, Li Fan, Wu Zhong, Li Rong. Emulation of Rocket Trajectory Based on a Six Degree of Freedom Model. Proceedings of SPIE Vol. 7128, 71282H, (2008) · 0277-786X/08/$18. DOI: 10.1117/12.806878

8. B. Milne, R. Carrese, A. Gardi, R. Sabatini. High-Fidelity Dynamics Modelling for the Design of a Sounding Rocket. Journal Designs. Vol. 7 (1), 32, 2023. https://doi.org/10.3390/designs7010032.

9. D. Cieśliński, R. Dziczkaniec, J. Kierski, C. Szczepański, M. Welcer. Simulation Model for Hardware-in-the-Loop Tests of a Rocket Control System**.** *Sensors*, vol. 25(13), 4083, 2025. https://doi.org/10.3390/s25134083.

10. N. Sahbon, M. Welcer. Comparison of Two Aerodynamic Models for Projectile Trajectory Simulation. Journal Aerospace. Vol 11(3), 189, 2024. https://doi.org/10.3390/aerospace11030189.

PSO-OPTIMIZED PID CONTROL IN GAS CONDENSATE ABSORPTION

G.Z. Ziyatbekova[1,3], A.M. Akhmetova[1], T.K. Zordunov[2], M.M. Sydykova[3], O.A. Sagyntay[1]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Kazakh National Agrarian Research University, Almaty, Kazakhstan
[3]Almaty Technological University, Almaty, Kazakhstan
E– mail: zordunovtimur@gmail.com

Abstract. *A detailed study of the automatic control system of a gas condensate absorber at the petrochemical plant is presented. The work spans the entire engineering process, from process description, mathematical modeling to synthesis of PID controller and intelligent parameter optimization applied to Particle Swarm Optimization (PSO). The mass balance equations of the absorber were used to develop a second-order transfer function and the stability of the transfer function was then established analytically using the Lyapunov, Nyquist, and Hurwitz criteria. This closed loop system was found to be stable but with poor overshoot (48.3%) and settling time (1.41 s) giving an incentive to design a PID controller. The Ziegler–Nichols (ZN), Water Cycle Algorithm (WCA), and Particle Swarm Optimization (PSO) tuning methods were tested and compared. Transient performance based on simulation in MATLAB/Simulink was conducted, resulting in the best combination of settling time and overshoot among the three algorithms, which were 0.41 s and 0.48 % for PSO-PID and 2.73 s and 0 % for ZN-PID. The optimized PSO-PID parameters are Kp = 0.3982, Ki = 9.8801, Kd = 0.1041. The study shows that the classical Ziegler–Nichols tuning is significantly outperformed by bio-inspired optimization techniques in the context of industrial absorber control and offers a replicable methodology for applications on other automation problems in the petrochemical industry.*

Keywords: *PID controller, gas condensate absorber, particle swarm optimization, mathematical modeling, transfer function.*

Introduction. For safe, efficient handling of complicated petrochemical operations, industrial automation is an integral part of the picture. Acid gas removal is typically accomplished with amine solutions in gas condensate processing plants which employ critical equipment for the removal of acid components from natural and associated gas streams, primarily hydrogen sulfide (H_2S) and carbon dioxide (CO_2) [1-3]. If the absorber level or pressure deviates from the set point, the product quality may be affected, equipment may be corroded or dangerous conditions may occur. Thus, the design of accurate and reliable automatic control systems for absorbers is current engineering research.

Proportional-Integral-Derivative (PID) controllers are still the most common type of feedback controller in process industries, which have been traditionally used and are proven to be simple and effective [8-10]. The classical Ziegler–Nichols (ZN) method, which was historically predominant, is however well known to yield controllers with too high an overshoot and settling times too long for processes with complex dynamics [4-7]. A suitable alternative is now available: Particle Swarm Optimization (PSO) [5] was proposed by Kennedy and Eberhart and has been successfully applied to many industrial optimal PID tuning applications.

The absorber system at the Pavlodar Petrochemical Plant is actually one of those really challenging control problems: A second order dynamic system with very much overshoot. This paper will discuss the entire engineering process of the mathematical model derivation, including the process of tuning the PID parameters using PSO, and comparing the performance with four other PID tuning techniques.

Materials and Methods. The control object is a vertical cylindrical absorber used for amine gas sweetening at a petrochemical plant. The absorber operates at pressures of 500–1500

kPa and temperatures of 30–70°C, with amine concentrations between 15 and 50 wt%. The controlled variable is the liquid level inside the absorber column, manipulated via a Level Control Valve (LCV).

Mathematical Model of the Absorber: Based on the geometric and hydrodynamic properties of the absorber, the liquid volume V as a function of level h follows the cylindrical cross-section relation. Under normal operating conditions (between high-level alarm LAH and low-level alarm LAL), the relationship is linearized:

$$V(h) = A \times L \times h(t), \ \ where \ \ A \approx \pi r^2$$

Applying the principle of mass balance to the liquid volume inside the absorber yields the dynamic equation:

$$A \times L \times \ dh(t)/dt \ = \ Q_{in}(t) \ - \ Q_{out}(t) \tag{1}$$

where Q_in(t) is the inflow rate and Q_out(t) is the outflow through the LCV, governed by:

$$Q_{out} = C_v \times f(u) \times \sqrt{\frac{\Delta P_{out}}{p_w}} \tag{2}$$

The valve coefficient Cv was estimated using the least-squares method on operational data. After linearizing around the nominal operating point and inserting system-specific parameters, the nonlinear model reduces to:

$$47.55 \times \frac{d(\Delta h)}{dt} = Q_{in}(t) - 1.81 \times \Delta h(t) - 10.82 \times \Delta u(t) \tag{3}$$

Taking the Laplace transform and normalizing, the closed-loop transfer function of the plant is obtained as:

$$G(s) = \frac{5.12s + 92.2}{s^2 + 5.16s + 92.2} \tag{4}$$

This is a second-order system with a numerator zero. The characteristic equation is, with complex conjugate roots $s_{1,2} = -2.582 \pm j9.252$. Since all real parts are strictly negative, the Lyapunov first method confirms closed-loop stability.

Stability Analysis:

Three stability criteria were applied to the derived transfer function:

Lyapunov (first method): All roots of the characteristic polynomial have negative real parts $(Re\,(s_{1,2}) = -2.582 < 0)$ ·The system is asymptotically stable.

Hurwitz criterion: For the polynomial $s^2 + 5.16s + 92.16$, the Hurwitz determinants $H_1 = 5.16 > 0 \, and \, H_2 = 5.16 \times 92.16 > 0$confirm stability.

Nyquist criterion: The Nyquist plot of the open-loop transfer function does not encircle the (−1, j0) point, confirming that the closed-loop system is stable.

Although the system is stable, open-loop step-response simulation in MATLAB/Simulink reveals a settling time of $T_{set} = 1.41 \ s$, overshoot $P_{ov} = 48.3\%$, and a steady-state error of $e\_ss = 0.5$. These values fail to meet industrial process control requirements, motivating PID synthesis.

PID Controller Synthesis: A standard PID controller with the control law:

$$u(t) = Kp \times e(t) + Ki \times \int e(t)dt + Kd \times de(t)/dt \tag{5}$$

was synthesized for the absorber model. Initial parameters were derived using the Ziegler–Nichols ultimate gain method, yielding the critical gain $Ku = 0.7964\ and\ period\ Pu = 0.6364\ s$, from which [7]:

$$Kp = 0.6Ku = 0.4778, \quad Ti = \frac{Pu}{2} = 0.3182, \quad Td = \frac{Pu}{8} = 0.0795 \tag{6}$$

These initial ZN parameters served as a baseline for comparing the subsequent optimization-based tuning methods.

PSO Optimization of PID Parameters: To overcome the limitations of ZN tuning, the Particle Swarm Optimization algorithm was applied to minimize the Integral Absolute Error (IAE) cost function [5]:

$$f(x) = \int_0^{\infty} |e(t)|\, dt \rightarrow min \tag{7}$$

In the PSO algorithm, each particle i maintains a position vector $X_i^t = (Kp,\ Ki,\ Kd)$ and a velocity vector V_i^t·The update equations at each iteration t are:

$$V_{ij}^{t+1} = w \cdot V_{ij}^t + c_1 r_1 (pbest_{ij} - X_{ij}^t) + c_2 r_2 (gbest_j - X_{ij}) \tag{8}$$

$$X_{ij}^{t+1} = X_{ij}^t + V_{ij}^{t+1} \tag{9}$$

where *w* is the inertia weight, c_1 and c_2 are cognitive and social acceleration coefficients, and r_1, r_2 are random numbers in [0,1]. The swarm converges to the global best position in the (Kp, Ki, Kd) parameter space, which minimizes IAE for the closed-loop system.

Three PID tuning methods were benchmarked: the classical Ziegler–Nichols (ZN) method, Particle Swarm Optimization (PSO), and the Water Cycle Algorithm (WCA).

Results and Discussion. The step responses of the three tuning methods were simulated in MATLAB/Simulink. Table 1 summarizes the key transient performance metrics [10].

Table 1. Comparative performance of PID tuning methods

Parameter	ZN-PID	PSO-PID	WCA-PID
Rise Time (s)	1.10	0.288	0.271
Settling Time (s)	2.731	0.414	0.609
Overshoot (%)	0.00	0.48	1.53
Peak Time (s)	14.51	1.15	0.79
Transient Time (s)	9.126	0.817	0.705
Kp	1.2	0.3982	0.5000
Ki	4	9.8801	10.000
Kd	0.05	0.1041	0.1000

The results reveal several important findings. The ZN-PID controller, while achieving zero overshoot, exhibits a dramatically long settling time of 2.731 s and a transient duration of 9.126 s — both unacceptable for real-time industrial control where rapid response to setpoint changes is critical. This sluggishness is a known limitation of the Ziegler–Nichols method when applied to systems with complex poles. This response profile is illustrated in Figure 1 below.

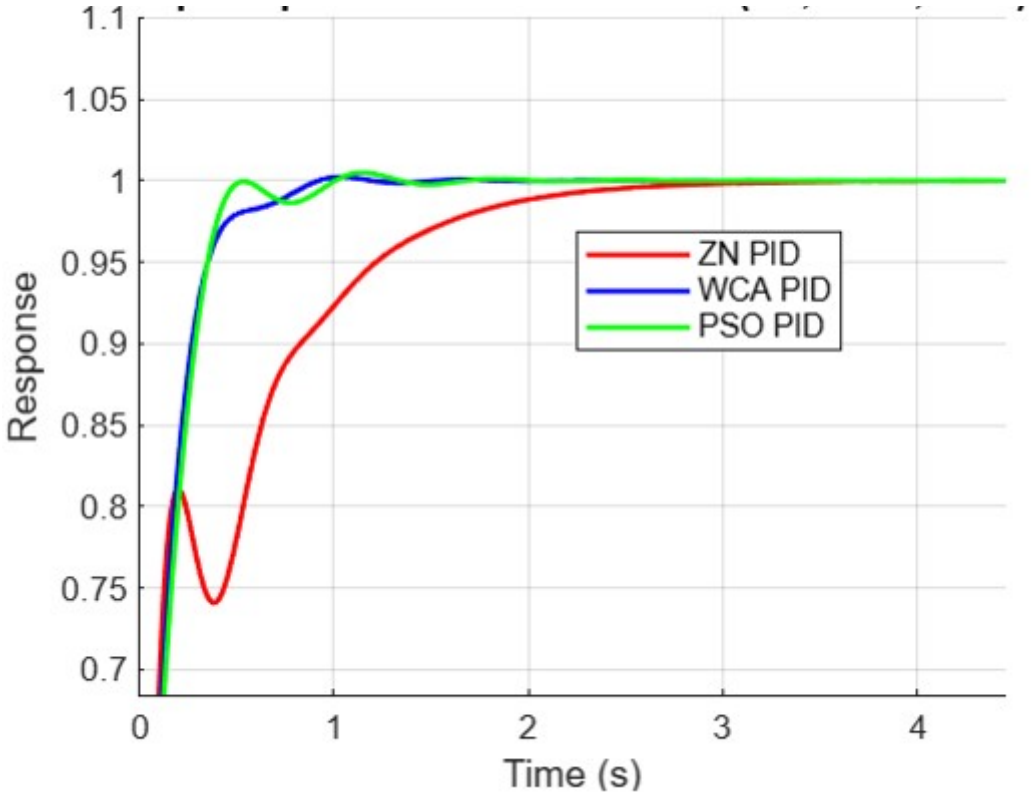


Figure 1. Step response of PID controllers

PSO-PID achieved the most favorable balance of performance metrics. Its settling time of 0.414 s represents a 6.6-fold improvement over ZN-PID, while maintaining overshoot below 0.5%. The high integral gain $Ki = 9.8801$reflects the algorithm's emphasis on eliminating steady-state error, consistent with the IAE minimization objective. WCA-PID showed comparable transient time but higher overshoot (1.53%), suggesting greater sensitivity to model uncertainties.

Table 2 presents the system quality metrics of the uncompensated closed-loop system before PID design.

Table 2. Closed-loop system quality metrics (before PID)

Parameter	Direct assessment	Indirect assessment
Settling time T_set (s)	1.41	1.55
Oscillation frequency ω (rad/s)	48.3	41.0
Degree of oscillation μ (%)	9.2	9.2
Number of oscillations n	16.0	3.58
Transient Time (s)	2	2
Rise time T_r (s)	0.11	—
Steady-state error e_ss	0.5	0

The system poles $s_{1,2} = -2.58 \pm j9.25 \; and \; s_3 = -18$indicate a dominant second-order response with moderate damping. The degree of oscillation $\mu = 3.58\%$ confirms

acceptable oscillatory behavior, but the 48.3% overshoot makes the uncompensated system unsuitable for direct application.

The PSO algorithm's strength lies in its ability to explore the three-dimensional parameter space (Kp, Ki, Kd) without requiring gradient information, making it robust to non-convex optimization landscapes typical of PID tuning. By minimizing the IAE objective, PSO naturally balances competing objectives of fast response and low steady-state error — a trade-off that classical frequency-domain methods cannot optimize simultaneously.

Conclusions. The main contribution of this study was the complete design and comparative analysis of an automated control system for a gas condensate absorber, which includes optimization of the PID parameters using PSO. The following conclusions are drawn: A second-order mathematical model of the absorber was derived from mass balance principles. Stability was confirmed via Lyapunov, Hurwitz, and Nyquist criteria, though the uncompensated system exhibited 48.3% overshoot and a settling time of 1.41 s. PSO-PID tuning reduced settling time to 0.414 s (6.6× improvement over ZN-PID) and overshoot to 0.48%, with optimized parameters $Kp = 0.3982,\ Ki = 9.8801,\ Kd = 0.1041$. Among the three methods compared (ZN, PSO, WCA), PSO-PID demonstrated the best overall transient performance, while ZN-PID showed the slowest response despite zero overshoot. The PSO algorithm's gradient-free, swarm-based search mechanism is particularly well-suited to PID tuning in petrochemical process control, where the optimization landscape is non-convex and system models contain inherent uncertainties. The results could be used in general to promote wider use of the bio-inspired optimization techniques for the tuning of industrial controllers, specifically in oil and gas production where both stability and speed and precision are needed. Further research should focus on testing robustness in the presence of parameter uncertainty, and experimental testing with physical absorber hardware.

References

1. World Bank (2023). Natural Gas Markets and Their Impacts on Energy Security // World Bank Energy Report. - https://www.worldbank.org/en/topic/naturalgas (09.09.2026).

2. Stenerson J. (2021). Industrial Automation and Process Control. Prentice Hall. - https://www.pearson.com (09.09.2026).

3. Johnson R. (2020). Instrumentation and Control Systems. McGraw Hill. - https://www.mheducation.com (09.09.2026).

4. Pinto S.B., Gertler J. (2021). Control and Automation Engineering. Springer.

5. Kennedy J., Eberhart R. (1995). Particle swarm optimization // Proceedings of ICNN. Vol. 4. Pp. 1942–1948. IEEE. DOI: 10.1109/ICNN.1995.488968.

6. Gaing Z.L. (2004). A particle swarm optimization approach for optimum design of PID controller in AVR system // IEEE Transactions on Energy Conversion. Vol. 19. №. 2. Pp. 384–391.

7. Ziegler J.G., Nichols N.B. (1942). Optimum settings for automatic controllers // Transactions of ASME. Vol. 64. Pp. 759–768.

8. Åström K.J., Hägglund T. (2006). Advanced PID Control. ISA.

9. Esmaili P. et al. (2023). PSO-based PID tuning for nonlinear industrial processes: a comparative study // Journal of Process Control. Vol. 123. Pp. 45–58. DOI: 10.1016/j.jprocont.2023.02.005.

10. MathWorks (2024). MATLAB Control System Toolbox Documentation. URL: https://www.mathworks.com/help/control.

Entropy-Guided Adaptive Scaffolding in Mathematics Education: A Bayesian Decision Model

V. Duran[1] and E. Hazar[1]

Iğdır University, Türkiye

Abstract. Adaptive support in mathematics is a decision under uncertainty: the teacher must infer which of several knowledge states best explains a learner's work and then choose a scaffold that does not replace the learner's reasoning. This conceptual paper formulates Entropy-Guided Adaptive Scaffolding (EGAS) as a constrained Bayesian decision model in which a belief over mutually exclusive knowledge states is revised from evidence, diagnostic uncertainty is measured by Shannon entropy, and candidate supports are scored by normalized expected information gain, expected learning gain, cognitive cost, and dependency risk. Three formal requirements are established. First, a purely inferential model has an expected learning gain that is identically zero for every action, so an explicit instructional transition kernel is necessary. Second, information gain must be normalized before it is combined with dimensionless pedagogical criteria. Third, low entropy is not evidence of competence, so fading must be conditioned jointly on concentration, expected mastery, calibration, and independent performance. A worked quadratic-factorization example illustrates the computations and a stability margin that states when the model should decline to issue a single recommendation. All numerical values are illustrative design inputs, not empirical results.

Keywords: Mathematics education · Adaptive scaffolding · Bayesian inference · Shannon entropy · Expected information gain · Knowledge tracing

1 Introduction

A mathematics teacher rarely observes knowledge directly. The available evidence is partial: a selected answer, an algebraic line, a hesitation, a hint request, a confidence judgement. Identical correctness can conceal different knowledge states—a conceptual misconception, a procedural slip, a missing verification habit, or excessive task load—so adaptive support must decide both what is likely and what to do next.

Intelligent tutoring research formalizes the inferential half of this problem through learner modelling and knowledge tracing [3–5], and decision-theoretic

tutoring extends it to action selection under partial observability [14, 15]. Prediction alone, however, does not determine an educationally appropriate intervention. Scaffolding theory requires contingency, gradual fading, and transfer of responsibility to the learner [6, 7]; cognitive load theory cautions that a mathematically informative prompt may still be poorly timed [8]; and formative feedback should modify thinking rather than supply the answer [9].

This paper specifies Entropy-Guided Adaptive Scaffolding (EGAS) and, in doing so, corrects three defects that an entropy-plus-utility formulation acquires if it is written down informally: an expected learning-gain term that is provably zero whenever beliefs are revised by Bayes' rule alone; a utility function that adds bits to dimensionless quantities; and a fading rule that treats low entropy as evidence of competence. The contribution is a formally consistent, auditable mechanism, not a claim of effectiveness: the model is analytical and reports no learner experiment.

2 Educational and Mathematical Foundations

Shannon entropy quantifies the dispersion of a probability distribution [1], and Lindley showed how an experiment may be valued by the information it is expected to provide [2]. A diagnostic question is an experiment in this restricted decision-theoretic sense: it is useful when its possible responses are expected to discriminate among plausible explanations of a learner's work. This interpretation does not reduce a learner to a number; it makes the assumptions behind a support decision visible and revisable.

Educationally, a scaffold is temporary support that enables activity the learner cannot yet perform independently [6], adjusted to evidence and faded as control transfers to the learner [7]. Mathematical competence includes reasoning, representation, communication, and verification rather than answer accuracy alone [11], so admissible evidence includes explanations, error localization, and transfer, and adaptation targets knowledge components in the sense of the Knowledge–Learning–Instruction framework [12]. Structurally, EGAS is a deliberately small partially observable decision process [14] of the kind applied to teaching by Rafferty et al. [15]; the state space is kept interpretable so that a teacher can inspect why an action was selected. EGAS operationalizes contingency but does not automate teacher judgement: states, likelihoods, mastery values, and risk limits must be specified for a curriculum and reviewed by educators.

3 The EGAS Decision Model

3.1 Belief State and Measurement Update

Let Z be a set of K mutually exclusive and exhaustive knowledge states for one curricular target, with z_0 denoting the secure state. Exhaustiveness matters: it is what makes the normalization in Eq. (1) legitimate. The belief p_t satisfies

$p_{t,k} \geq 0$ and $\sum_k p_{t,k} = 1$. After action a produces evidence e, the measurement (filtering) update is

$$\widetilde{p}_{t,k}(e,a) = \frac{P(e \mid z_k, a)p_{t,k}}{\sum_j P(e \mid z_j, a)p_{t,j}}, \qquad P(e \mid a) = \sum_k P(e \mid z_k, a)p_{t,k}. \tag{1}$$

The update is defined whenever $P(e \mid a) > 0$. Likelihoods may be estimated from pilot data, elicited from experts, or learned from validated logs; they must not be invented and then reported as measurements.

3.2 Uncertainty and Expected Information Gain

Write

$$H(p) = -\sum_k p_k \log_2 p_k, \qquad h_t = \frac{H(p_t)}{\log_2 K} \in [0,1], \quad K \geq 2, \tag{2}$$

with the convention $0 \log_2 0 = 0$. For an action a with evidence set $\mathcal{E}_a$, the expected information gain is the reduction in entropy produced by the measurement step, which is the mutual information between state and evidence under the joint distribution that a induces:

$$IG(t,a) = H(p_t) - \sum_{e \in \mathcal{E}_a} P(e \mid a) H(\widetilde{p}_t(e,a)) = I_a(Z_t; E). \tag{3}$$

Proposition 1. $IG(t,a) \geq 0$ for every a, with equality if and only if $P(e \mid z_k, a)$ is the same for every k with $p_{t,k} > 0$; states currently held impossible carry no information. This follows from the concavity of H and Jensen's inequality. Two consequences are used below: an uninformative action can never be rewarded by the first criterion, and $IG \leq H(p_t) \leq \log_2 K$, which is what makes the normalization in Sect. 3.4 well defined. Low entropy, however, is not in itself desirable—it may express confident belief in a persistent misconception—and this is why entropy alone never triggers fading in Sect. 3.5.

3.3 Instructional Effect and Learning Gain

Let $M : Z \to [0,1]$ be a curriculum-defined mastery function with $M(z_0) = 1$, and let $m_t = \sum_k M(z_k)p_{t,k}$ be expected mastery. It is tempting to score an instructional action by the expected change in m_t. Under Eq. (1) alone this is vacuous.

Proposition 2. If the latent state is static, the evidence set $\mathcal{E}_a$ is exhaustive, and beliefs are revised only by Eq. (1), then for every action a,

$$\mathbb{E}[m_{t+1} \mid a] = \sum_e P(e \mid a) \sum_k M(z_k)\widetilde{p}_{t,k}(e,a) \tag{4}$$

$$= \sum_k M(z_k)p_{t,k} \sum_e P(e \mid z_k, a) = m_t. \tag{5}$$

Thus the expected learning gain is identically zero: the posterior mean is a martingale with respect to the prior. Bayesian revision changes what we believe about the learner, not what the learner knows. An entropy-guided model that also claims to value learning therefore requires an explicit instructional effect. We add a transition kernel $T(z_j \mid z_k, a)$, the probability that action a moves a learner in state z_k to z_j, and complete each cycle as prediction after correction:

$$p_{t+1,j}(e,a) = \sum_k T(z_j \mid z_k, a)\widetilde{p}_{t,k}(e,a), \qquad m^T(z_k,a) = \sum_j T(z_j \mid z_k, a)M(z_j), \tag{6}$$

$$G(t,a) = \mathbb{E}[m_{t+1} \mid a] - m_t = \sum_k p_{t,k}\left[m^T(z_k,a) - M(z_k)\right]. \tag{7}$$

$G(t,a)$ is now generally nonzero and is positive exactly when a is expected to move probability mass toward states of higher mastery. Note that IG is computed on the measurement step only. The transition step in Eq. (6) may either raise or lower entropy—a kernel that concentrates mass on z_0 lowers it—and in either direction the resulting spread describes a different random variable, the post-instruction state; folding it into Eq. (3) would confound what was learned about the learner with what the learner learned.

3.4 Commensurable Utility and Constrained Choice

IG is measured in bits and ranges over $[0, \log_2 K]$, whereas $G \in [-1,1]$ and the cognitive cost $C_t(a) \in [0,1]$ and dependency risk $D_t(a) \in [0,1]$ are dimensionless. Adding them directly is not meaningful. We therefore use the normalized gain $ig(t,a) = IG(t,a)/\log_2 K \in [0,1]$ and constrain the weights to the simplex, which also removes the scale indeterminacy of a linear utility:

$$U(t,a) = \alpha\, ig(t,a) + \beta\, G(t,a) - \lambda C_t(a) - \mu D_t(a), \quad \alpha+\beta+\lambda+\mu = 1, \quad \alpha,\beta,\lambda,\mu \geq 0. \tag{8}$$

Safety, access, and the current support level are applied before maximization. Let $d(a) \in [0,1]$ be the directiveness of a and $s_t \in [0,1]$ the support level maintained by Eqs. (12)–(13); coupling them is what prevents the utility rule and the fading rule from contradicting each other:

$$\mathcal{A}_{\text{safe}}(t) = \{a \in \mathcal{A}(t) : d(a) \leq s_t,\ C_t(a) \leq C_{\max},\ D_t(a) \leq D_{\max}\}, \tag{9}$$

$$a^*(t) = \arg\max_{a \in \mathcal{A}_{\text{safe}}(t)} U(t,a). \tag{10}$$

Ties are broken toward smaller D. If $\mathcal{A}_{\text{safe}}(t)$ is empty the model selects nothing and refers the decision to the teacher; an unconstrained argmax over a possibly empty set is undefined and must not be left implicit.

3.5 Calibration and Contingent Fading

For reported confidence $c_t \in [0,1]$ and correctness $y_t \in \{0,1\}$, EGAS tracks a smoothed Brier score [13]:

$$B_t = (c_t - y_t)^2, \qquad \overline{B}_t = \rho\overline{B}_{t-1} + (1-\rho)B_t, \qquad \overline{B}_0 = B_0, \quad 0 \leq \rho < 1. \tag{11}$$

B_t is a strictly proper scoring rule and $\overline{B}_t$ its exponentially weighted average, low only when confidence is both accurate and honest; it is not a pure calibration index, since it also absorbs resolution, and should be reported as a Brier score rather than as calibration error. Fading is then conditioned on four signals rather than on entropy alone, and support is re-escalated through a wider band so that the rule does not oscillate:

$$s_{t+1} = \max(0, s_t - \delta) \quad \text{if} \quad h_t \le \tau_h,\ m_t \ge \tau_m,\ \overline{B}_t \le \tau_B,\ q_t \ge \tau_q, \tag{12}$$

$$s_{t+1} = \min(1, s_t + \delta) \quad \text{if} \quad h_t \ge \tau_h^{\uparrow} \text{ or } q_t \le \tau_q^{\downarrow}; \qquad s_{t+1} = s_t \text{ otherwise}, \tag{13}$$

with $q_t \in [0, 1]$ the independently demonstrated success estimate, $\delta > 0$, and $\tau_h^{\uparrow} > \tau_h$, $\tau_q^{\downarrow} < \tau_q$. The condition $m_t \ge \tau_m$ is what separates "the model is certain" from "the learner is competent": without it, a confidently diagnosed misconception—a low-entropy, low-mastery belief—would trigger the removal of support.

4 Worked Mathematical Example

Consider $x^2 - 5x + 6 = 0$. A learner writes $(x - 2)(x + 3) = 0$ and reports $x = 2, -3$. Since $(x - 2)(x + 3) = x^2 + x - 6$ while $(x - 2)(x - 3) = x^2 - 5x + 6$, the sign choice violates both coefficient conditions at once—the constant term becomes -6 instead of 6 and the linear coefficient $+1$ instead of -5—although the magnitudes 2 and 3 are correct. The roots are nonetheless read correctly from the learner's own expression, so the failure lies in factor selection, not in applying the zero-product property.

Four exclusive states are defined: z_0, secure factorization with verification, $M = 1$; z_1, a systematic sign error in factor selection, with neither coefficient condition checked against the expansion, $M = 0.25$; z_2, correct procedure with an execution slip and no verification habit, $M = 0.70$; z_3, an unstable or partial factorization schema, $M = 0.10$. Conditioned on the observed error, $p_0 = (0, 0.40, 0.35, 0.25)$, giving $H(p_0) = 1.559$ bits, $h_0 = 1.559/\log_2 4 = 0.779$, and $m_0 = 0.370$; because h is normalized by the full state space, a belief that is almost uniform on a proper subset still reports $h < 1$. A zero prior on z_0 cannot be revived by Eq. (1), but the transition kernel can move mass into it—which is exactly the gap Proposition 2 identifies.

Take a_4, comparing $(x - 2)(x + 3)$ with $(x - 2)(x - 3)$, and let $e = 1$ mean that the learner expands both products, matches coefficients, and locates the mismatch without being given the answer. Likelihoods must respect the state definitions. Expansion is intact under both z_1 and z_2, so once the comparison is supplied both are likely to locate the mismatch—the slip state slightly more so, since no faulty rule can intrude on the expansion—whereas an unstable schema is much less likely. Take $P(e = 1 \mid z_1, a_4) = 0.70$, $P(e = 1 \mid z_2, a_4) = 0.85$, and $P(e = 1 \mid z_3, a_4) = 0.20$. Then $P(e = 1 \mid a_4) = 0.6275$, the posteriors are $(0, 0.446, 0.474, 0.080)$ and $(0, 0.322, 0.141, 0.537)$ with entropies 1.321 and 1.407 bits, their weighted mean is 1.353 bits, and $IG_0(a_4) = 0.206$ bits, $ig_0(a_4) =$

Table 1. Illustrative action scores under $\alpha = 0.45$, $\beta = 0.30$, $\lambda = 0.15$, $\mu = 0.10$.

Instructional action	IG (bits)	ig	G	C	D	U
a_1 Generic hint	0.060	0.030	0.25	0.15	0.30	0.0360
a_2 Substitution check	0.220	0.110	0.35	0.18	0.08	0.1195
a_3 Worked example	0.100	0.050	0.45	0.12	0.40	0.0995
a_4 Compare factorizations	0.206	0.103	0.4185	0.28	0.10	0.1199

0.103. The action separates z_1 from z_2 only weakly, so most of its value must come from the second criterion rather than the first. Expected mastery under the two posteriors is 0.451 and 0.233, and weighting these by 0.6275 and 0.3725 returns exactly $m_0 = 0.370$—Proposition 2 in the example.

For the instructional effect, let the rows of $T(\cdot \mid z_k, a_4)$ over (z_0, z_1, z_2, z_3) be $(0.70, 0.20, 0.05, 0.05)$ from z_1, $(0.80, 0.05, 0.10, 0.05)$ from z_2, and $(0.45, 0.15, 0.20, 0.20)$ from z_3; the likelihood and the kernel row attached to z_0 are immaterial here because $p_0(z_0) = 0$. Equation (6) gives $m^T = 0.7900, 0.8875, 0.6475$ respectively, so by Eq. (7), $G_0(a_4) = 0.7885 - 0.3700 = 0.4185$. Expected mastery therefore rises from 0.3700 to 0.7885 under this kernel: the comparison earns most of its utility instructionally rather than diagnostically, which is the reverse of what an entropy-first reading would predict.

Entries other than the a_4 row are illustrative design values, not observed student data. Equation (8) gives $U_0(a_4) = 0.45(0.103) + 0.30(0.4185) - 0.15(0.28) - 0.10(0.10) = 0.1199$, nominally the largest, but it exceeds $U_0(a_2) = 0.1195$ by 0.0004. The associated pedagogical act is mathematically substantive: expand both products, match coefficients against $x^2 - 5x + 6$, then verify the candidate roots by substitution. It diagnoses and teaches while leaving the reasoning with the learner—but on these inputs the model has no warrant to prefer it to a substitution check.

A ranking this close must not be reported as a decision. For actions a and b, the indifference condition is linear in the weight vector $w = (\alpha, \beta, \lambda, \mu)$: writing

$$\Delta(a, b) = \big(ig(a) - ig(b),\ G(a) - G(b),\ -[C(a) - C(b)],\ -[D(a) - D(b)]\big)$$

gives $U(a) - U(b) = w \cdot \Delta(a, b)$. Any admissible reweighting must keep the weights summing to one, so the relevant direction is not $\Delta(a, b)$ itself but its projection $\overline{\Delta}(a, b)$ onto the sum-zero subspace, that is $\Delta(a, b)$ minus its own mean. The smallest relative perturbation of w that reverses the ranking is then

$$\varepsilon(a, b) = \frac{|U(t, a) - U(t, b)|}{\|\overline{\Delta}(a, b)\|_2 \, \|w\|_2}. \tag{14}$$

Here $\varepsilon(a_4, a_1) = 0.572$, $\varepsilon(a_4, a_3) = 0.107$, but $\varepsilon(a_4, a_2) = 0.006$; for a_4 against a_1 the minimizing reweighting leaves the nonnegative orthant, so 0.572 is a lower bound and the feasible margin is larger. If $\overline{\Delta}(a, b) = 0$, the ranking is invariant to every admissible reweighting and ε is taken as infinite. Under a declared threshold $\varepsilon_{\min} = 0.10$, the model robustly rejects the generic hint and,

marginally, prefers comparison to a worked example, but it does not resolve the choice between comparing factorizations and requiring a substitution check: a reweighting far smaller than ordinary expert disagreement reverses it. EGAS should then report both actions with their component scores rather than conceal the alternative behind a single label. What this example delivers reliably is an exclusion, not a selection, and reporting ε converts the fragility of a linear utility from a hidden defect into an explicit output.

5 Implementation, Evaluation, and Limitations

Implementation begins with a curriculum map. Teachers define a small set of interpretable states for a topic, identify the evidence that distinguishes them, and order support actions by directiveness $d(a)$: request an explanation, ask for a boundary case, prompt a representation change, offer a targeted hint, compare cases, provide a partial worked step. The map also fixes $C_{\max}$, $D_{\max}$, the fading thresholds, and what counts as independent performance. Written mathematics remains the primary evidence; response time and click data are supplementary and are not neutral proxies for knowledge. Three assessment objects feed the model: mathematical accuracy, evidence quality (justification, representation, verification), and calibration, aligning adaptation with broad competence [11, 12] while the dependency penalty discourages unlimited hint escalation and feedback stays specific and actionable [9].

An adequate empirical test would compare EGAS-guided scaffolding with fixed-sequence hints and teacher-selected support in a cluster-randomized or crossover classroom design, with outcomes including near and far transfer, independent problem solving after supports are withdrawn, calibration, hint dependence, and cognitive load. Model evaluation should report posterior calibration, predictive log loss, action frequencies, subgroup error, and sensitivity to priors, kernels, and weights. Prior tutoring research offers comparison evidence [10], but EGAS effectiveness must be established directly rather than inferred from it. The transition kernel is the parameter most in need of empirical estimation, since it now carries the model's entire claim to instructional value.

Limitations follow from the formalization itself. Exhaustive, mutually exclusive states are a modelling idealization; real difficulties co-occur, and a misspecified partition produces precisely calculated but educationally invalid actions. Likelihood and kernel estimates may be sparse. A scalar utility cannot represent teacher values completely, which is why ε in Eq. (14) is reported rather than suppressed. Entropy can privilege diagnostic efficiency over dialogue. Accordingly EGAS is a decision aid with auditable assumptions, not an autonomous diagnostician, and privacy-preserving data collection, teacher override, and periodic bias review are design conditions rather than optional additions.

6 Conclusion

EGAS formulates adaptive scaffolding in mathematics as a constrained Bayesian decision problem in which beliefs are corrected by evidence and predicted through an instructional transition, actions are valued on a single normalized scale, and support fades only when concentration, expected mastery, calibration, and independent performance jointly warrant it. Its educational contribution is not the replacement of teacher judgement but the exposure of hidden assumptions: what the system believes, why a question is informative, what a hint may cost, when support should disappear, and—through the stability margin—when the model should decline to answer. Future work should estimate and validate the likelihoods and transition kernels in authentic classrooms and test whether the model improves transfer and learner autonomy.

References

1. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
2. Lindley, D.V.: On a measure of the information provided by an experiment. Ann. Math. Stat. **27**(4), 986–1005 (1956). https://doi.org/10.1214/aoms/1177728069
3. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model. User-Adapt. Interact. **4**, 253–278 (1994). https://doi.org/10.1007/BF01099821
4. Desmarais, M.C., Baker, R.S.J.d.: A review of recent advances in learner and skill modeling in intelligent learning environments. User Model. User-Adapt. Interact. **22**, 9–38 (2012). https://doi.org/10.1007/s11257-011-9106-8
5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems 28, pp. 505–513 (2015)
6. Wood, D., Bruner, J.S., Ross, G.: The role of tutoring in problem solving. J. Child Psychol. Psychiatry **17**(2), 89–100 (1976). https://doi.org/10.1111/j.1469-7610.1976.tb00381.x
7. van de Pol, J., Volman, M., Beishuizen, J.: Scaffolding in teacher–student interaction: A decade of research. Educ. Psychol. Rev. **22**, 271–296 (2010). https://doi.org/10.1007/s10648-010-9127-6
8. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cogn. Sci. **12**(2), 257–285 (1988). https://doi.org/10.1207/s15516709cog1202_4
9. Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008). https://doi.org/10.3102/0034654307313795
10. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011). https://doi.org/10.1080/00461520.2011.611369
11. Niss, M., Højgaard, T.: Mathematical competencies revisited. Educ. Stud. Math. **102**, 9–28 (2019). https://doi.org/10.1007/s10649-019-09903-9
12. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The Knowledge–Learning–Instruction framework: Bridging the science–practice chasm to enhance robust student learning. Cogn. Sci. **36**(5), 757–798 (2012). https://doi.org/10.1111/j.1551-6709.2012.01245.x

13. Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. **78**(1), 1–3 (1950). https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
14. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artif. Intell. **101**(1–2), 99–134 (1998). https://doi.org/10.1016/S0004-3702(98)00023-X
15. Rafferty, A.N., Brunskill, E., Griffiths, T.L., Shafto, P.: Faster teaching via POMDP planning. Cogn. Sci. **40**(6), 1290–1332 (2016). https://doi.org/10.1111/cogs.12290

The VERIFY Cycle: A Verification-First Pedagogy for Generative AI in Mathematical Modelling

V. Duran[1] and E. Hazar[1]

Iğdır University, Türkiye

Abstract. Generative artificial intelligence can produce rapid mathematical solutions, but fluent output may encourage answer acceptance rather than reasoning. This conceptual paper develops VERIFY, a six-stage verification-first cycle for AI-supported mathematical modelling: Verbalize, Estimate, Request, Inspect, Falsify, and Yield. The framework was derived through theory-guided synthesis of research on generative AI in mathematics education, modelling, creative reasoning, productive failure, comparison, and self-regulated learning. An illustrative cooling-model lesson shows how students test assumptions, units, representations, and boundary behavior before defending a conclusion. VERIFY treats AI as a fallible conjecture generator and makes the learner's verification trail the principal assessment object.

Keywords: Generative artificial intelligence · Mathematics education · Mathematical modelling · Verification · Self-regulated learning

1 Introduction

Generative artificial intelligence (GenAI) can explain procedures, propose models, produce code, and respond immediately to follow-up questions. Its educational value, however, depends on the activity it enables rather than on the apparent quality of its prose. UNESCO therefore frames GenAI use as a human-centred design problem that should preserve human agency, privacy, and pedagogical purpose [1]. Broader reviews similarly identify useful functions alongside practical risks, including inaccurate output, weak transparency, over-reliance, and uneven readiness among teachers and institutions [2].

These tensions are unusually important in mathematics education. Recent reviews report possibilities for feedback, problem generation, explanation, and personalized support, but they also warn that students may accept plausible errors or outsource core reasoning [3, 4]. Comparisons between AI- and teacher-generated problem-solving guidance show that linguistic fluency does not by itself guarantee mathematically or pedagogically sufficient guidance [5]. Research on AI-supported proving likewise indicates that students need explicit support to evaluate, rather than passively adopt, generated arguments [6].

The central educational problem is thus not whether GenAI can return an answer. It is whether classroom design keeps students responsible for mathematizing a situation, choosing representations, checking assumptions, and defending conclusions. This paper proposes the VERIFY cycle, a verification-first instructional framework for GenAI-supported mathematical modelling. Its contribution is a teachable sequence that positions AI as a fallible generator of conjectures and comparison cases while making the student's evidence trail—not the AI's final response—the object of assessment.

2 Conceptual Method and Design Principles

VERIFY was developed through a theory-guided integrative synthesis, not a claim of systematic review or intervention effectiveness. Four bodies of scholarship were connected because each addresses a different requirement of responsible AI-supported mathematics learning. Recent GenAI reviews establish the opportunity-risk landscape [2–4]; mathematical modelling research emphasizes movement between a real situation, a mathematical model, mathematical work, interpretation, and validation [7]; competence and reasoning research distinguishes active mathematical inquiry from the reproduction of familiar procedures [8, 9]; and educational psychology explains why prior generation, comparison, monitoring, and reflection can support learning [10–12]. Sources were limited to official guidance, peer-reviewed syntheses, direct empirical studies, and established theoretical works with verifiable publication records.

The synthesis produced four design principles. First, students make their own problem representation and estimate before seeing AI output. Second, AI is asked for alternatives and assumptions, not treated as an authority. Third, verification must use more than rereading; it should include mathematical substitution, units, boundary cases, representation switching, counterexamples, or computational checks. Fourth, the learning product records changes in the student's judgement. These principles turn potential AI error into a planned contrast while preventing error exposure from becoming the lesson's sole purpose.

3 The VERIFY Cycle

Figure 1 presents six recursive stages. A teacher can use the full cycle for a modelling investigation or compress it into a short routine for one problem. The return arrow is essential: when verification reveals a defective assumption or representation, students revise the problem framing rather than merely repair the last calculation.

V – Verbalize. Students restate the problem, identify quantities and relationships, specify what is unknown, and list assumptions. The teacher asks what is given, what is inferred, and what evidence would make a model acceptable. This stage prevents the AI prompt from becoming the first and only representation of the task.

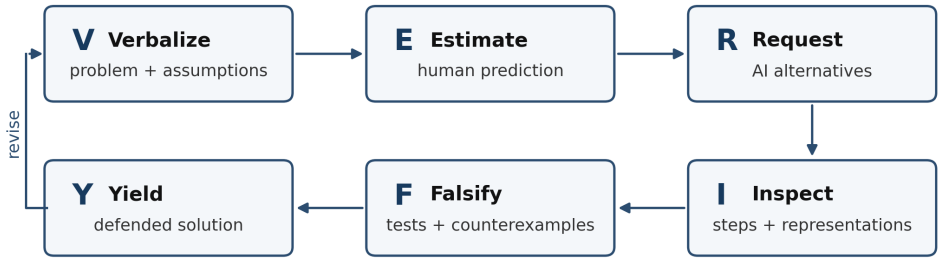


Fig. 1. The recursive VERIFY cycle for AI-supported mathematical modelling.

E – Estimate. Before using AI, students make an order-of-magnitude prediction, sketch expected behavior, or propose a candidate structure. The estimate need not be exact; it creates a reference point for monitoring. Productive-failure research supports the value of generating ideas before receiving a polished solution [10].

R – Request. Students ask GenAI for at least two possible approaches, the assumptions behind each, and conditions under which each approach could fail. Prompt quality matters, but the aim is diversity of inspectable claims rather than a more persuasive final answer. The prompt and response are retained as evidence.

I – Inspect. Students annotate the AI output line by line. They connect symbols to quantities, check transformations, compare representations, and mark unsupported transitions. Comparison research shows that contrasting methods can strengthen conceptual and procedural knowledge when learners attend to meaningful similarities and differences [11].

F – Falsify. Students actively try to break the proposed solution. They substitute results, test units and signs, examine extreme or boundary cases, construct counterexamples, graph behavior, or reproduce a calculation with a calculator or short program. The question changes from “Does this look right?” to “What test could show that it is wrong?”

Y – Yield. Students produce a defended solution that distinguishes accepted, revised, and rejected AI contributions. A brief reflection states what changed between the initial estimate and final model, which check was decisive, and what uncertainty remains. This stage closes the self-regulatory loop of planning, monitoring, evaluation, and adaptation [12].

4 Illustrative Lesson: Modelling the Cooling of a Drink

Consider an upper-secondary or introductory university task: A drink is 90 °C in a 22 °C room and is 64 °C after 10 minutes. When will it reach 45 °C?

The numbers are illustrative rather than research data. During Verbalize, students identify temperature, time, room temperature, and assumptions such as a stable environment. During Estimate, they predict a plausible interval and sketch a decreasing curve that approaches, but should not pass far below, room temperature.

At Request, GenAI is asked to propose both a linear model and Newton's exponential cooling model. During Inspect, students compare the consequences. A linear continuation gives

$$T(t) = 90 - 2.6t, \tag{1}$$

and predicts 45 °C at approximately 17.3 minutes, but it eventually predicts temperatures below room temperature. For the exponential model,

$$T(t) = 22 + 68\exp(-kt). \tag{2}$$

The 10-minute observation gives

$$k = -\frac{\ln(42/68)}{10} \approx 0.0482 \text{ min}^{-1}, \tag{3}$$

and therefore

$$t = \frac{\ln(68/23)}{k} \approx 22.5 \text{ min} \tag{4}$$

for 45 °C.

During Falsify, students check units, substitute $t = 0$ and $t = 10$, inspect long-run behavior, and compare both models against a later measurement supplied by the teacher. A graphing tool or short script may be used as an independent representation, but it does not replace the verbal interpretation. During Yield, students defend the exponential model within its assumptions, explain why the linear model can be locally useful but globally implausible, and record whether the AI omitted any assumption or check. The instructional value comes from comparing and validating models, not from reproducing the numerical answer.

5 Assessment and Implementation

A verification-first lesson requires process-aligned assessment. A compact analytic rubric can score five dimensions from 0 to 2: quality of problem framing, independence and reasonableness of the initial estimate, depth of method comparison, diversity and correctness of verification tests, and strength of the final justification. The assessable evidence consists of the initial representation, estimate, prompt-response record, annotated checks, and final explanation. A correct answer with no verification trail should therefore receive less credit than a well-documented analysis that detects and repairs an error.

Implementation can begin with three teacher controls. Access to GenAI is delayed until an independent estimate exists; every AI-supported solution must undergo at least two different checks; and the final submission must disclose

what was retained, changed, or rejected. Teacher questioning remains important, especially for novices who may not know which boundary cases or representations are diagnostic. The cycle is not a substitute for direct instruction. It is a structure for deciding when explanation, worked examples, peer comparison, or targeted feedback should enter the lesson.

Responsible implementation also requires age-appropriate access, protection of personal data, and equitable participation [1, 2]. Teachers can use institutionally approved accounts, remove identifying information from prompts, provide a non-AI comparison case, and organize paired verification when device access is uneven. Model name, version, date, and the exact prompt should be recorded because outputs can change. These safeguards make the lesson reproducible enough for review without portraying a stochastic system as a stable source.

6 Discussion and Research Agenda

VERIFY contributes an educational design response to a problem identified across the GenAI literature: fluent assistance can support learning while simultaneously concealing error and weakening student agency [2–6]. The framework changes the unit of success from answer production to an auditable chain of judgement. It also connects modelling competence with self-regulated learning: students define standards, compare a generated product with those standards, take corrective action, and explain the adaptation [7, 8, 12].

The framework remains conceptual and should not be interpreted as evidence of improved achievement. Future studies could compare unrestricted GenAI use, VERIFY-guided use, and non-AI instruction on conceptual transfer, error detection, calibration of confidence, and modelling quality. Process data could examine whether learners actually revisit assumptions after falsification and which verification strategies predict successful transfer. Studies should also test how prior knowledge, task complexity, and teacher scaffolding alter the cycle; novices may need checklists and worked examples, whereas advanced learners may benefit from deliberately subtle AI errors.

7 Conclusion

GenAI should not become the endpoint of mathematical activity. In the VERIFY cycle, its output is provisional material for modelling, comparison, falsification, and reflection. The resulting pedagogy keeps efficiency without surrendering epistemic responsibility: students may use AI to widen the solution space, but they remain accountable for assumptions, tests, and conclusions.

Data Statement No empirical data were generated for this conceptual paper.

AI Disclosure The authors utilized GPT-5.6 Sol exclusively for grammar correction, text formatting, and language editing.

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Miao, F., Holmes, W.: Guidance for Generative AI in Education and Research. UNESCO (2023). https://doi.org/10.54675/EWZM9535
2. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., Gašević, D.: Practical and ethical challenges of large language models in education: A systematic scoping review. Br. J. Educ. Technol. **55**(1), 90–112 (2024). https://doi.org/10.1111/bjet.13370
3. Pepin, B., Buchholtz, N., Salinas-Hernández, U.: A scoping survey of ChatGPT in mathematics education. Digit. Exp. Math. Educ. **11**(1), 9–41 (2025). https://doi.org/10.1007/s40751-025-00172-1
4. Turmuzi, M., Azmi, S., Kertiyani, N.M.I.: ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. Teach. Teach. Educ. **170**, 105286 (2026). https://doi.org/10.1016/j.tate.2025.105286
5. Jia, J., Wang, T., Zhang, Y., Wang, G.: The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers. Asia Pac. J. Educ. **44**(1), 8–28 (2024). https://doi.org/10.1080/02188791.2023.2286920
6. Yoon, H., Hwang, J., Lee, K., Roh, K.H., Kwon, O.N.: Students' use of generative artificial intelligence for proving mathematical statements. ZDM Math. Educ. **56**(7), 1531–1551 (2024). https://doi.org/10.1007/s11858-024-01629-0
7. Blum, W., Leiß, D.: How do students and teachers deal with modelling problems? In: Haines, C., Galbraith, P., Blum, W., Khan, S. (eds.) Mathematical Modelling: Education, Engineering and Economics, pp. 222–231. Horwood Publishing (2007). https://doi.org/10.1533/9780857099419.5.221
8. Niss, M., Højgaard, T.: Mathematical competencies revisited. Educ. Stud. Math. **102**(1), 9–28 (2019). https://doi.org/10.1007/s10649-019-09903-9
9. Lithner, J.: A research framework for creative and imitative reasoning. Educ. Stud. Math. **67**(3), 255–276 (2008). https://doi.org/10.1007/s10649-007-9104-2
10. Kapur, M.: Productive failure. Cogn. Instr. **26**(3), 379–424 (2008). https://doi.org/10.1080/07370000802212669
11. Rittle-Johnson, B., Star, J.R.: Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. J. Educ. Psychol. **99**(3), 561–574 (2007). https://doi.org/10.1037/0022-0663.99.3.561
12. Panadero, E.: A review of self-regulated learning: Six models and four directions for research. Front. Psychol. **8**, 422 (2017). https://doi.org/10.3389/fpsyg.2017.00422

УДК 004.94:004.946

INTEGRATING WEB-BASED AUGMENTED REALITY INTO INDUSTRIAL DASHBOARDS: A CASE STUDY ON REAL-TIME FACTORY EQUIPMENT MONITORING

Nurgazy T.[1], Amirkhanova G.[1], Amirkhanov B.[1], Rustemov A.[1], Liu H.[2]
e-mail: nurgazytomiris@gmail.com
[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[2]*Peking University, Beijing, China*

Abstract. *The rapid advancement of Industry 4.0 requires modern approaches to monitoring industrial equipment. While conventional 2D dashboards provide comprehensive data analytics, they lack spatial context on the factory floor. This paper presents a framework for embedding Web-based Augmented Reality (WebAR) directly into strict corporate monitoring platforms, specifically Grafana. We detail a case study of a bakery factory where real-time sensor data from an InfluxDB time-series database is visualized as interactive 3D holograms over physical equipment using AR.js and A-Frame. The study addresses major technical challenges: establishing secure HTTPS connections via zero-trust tunnels to bypass local browser WebRTC restrictions, overcoming Grafana's rigid HTML sanitization policies within a Kubernetes environment, and optimizing the mobile AR user experience through quaternion-based billboard effects for spatial UI stabilization. The resulting architecture offers a zero-install, scalable mobile AR solution equipped with edge-computing algorithms for immediate on-site diagnostics and anomaly detection, such as phase imbalance and operational overloads.*

Keywords: *augmented reality; WebAR; digital twin; Industrial IoT; Grafana; Kubernetes; real-time monitoring.*

1. Introduction

The transition toward Industry 4.0 and 5.0 positions Digital Twins (DTs) at the center of industrial modernization [1]. For managing massive IIoT telemetry, cloud-native architectures like Kubernetes and Grafana have become industry standards [2], often paired with InfluxDB to process high-frequency data for real-time decision-making [3]. However, raw data analytics is insufficient without spatial 3D visualization [4]. Relying on remote 2D dashboards creates a cognitive gap, forcing maintenance engineers to constantly switch context between abstract metrics and physical machinery. Augmented Reality (AR) bridges this divide by superimposing real-time holograms directly onto machinery, accelerating fault detection [5]. To bypass the deployment barriers of native apps, zero-install Web-based AR (WebAR) leverages WebGL and WebRTC for browser-based diagnostics [6]. Despite its potential, embedding WebAR into rigid platforms like Grafana presents major challenges. Corporate Kubernetes environments employ strict HTML sanitization that blocks AR scripts [7], WebRTC mandates complicate integration with unencrypted factory networks, and rendering dynamic panels requires complex quaternion mathematics for viewport stabilization [8].

This paper presents a practical approach for directly integrating marker-based WebAR into industrial Grafana dashboards. By bypassing Kubernetes-level security constraints, securing local-network WebRTC via zero-trust tunnels, and implementing a mathematical “billboard effect”, we deliver a reliable mobile AR tool for immediate on-site anomaly detection, such as phase imbalance.

2. System architecture

To enable real-time augmented reality monitoring within a strict enterprise environment, a robust four-tier architecture was developed [9]. As depicted in Fig. 1, the system integrates physical hardware, edge data processing, cloud-based visualization, and mobile AR rendering into a cohesive pipeline.

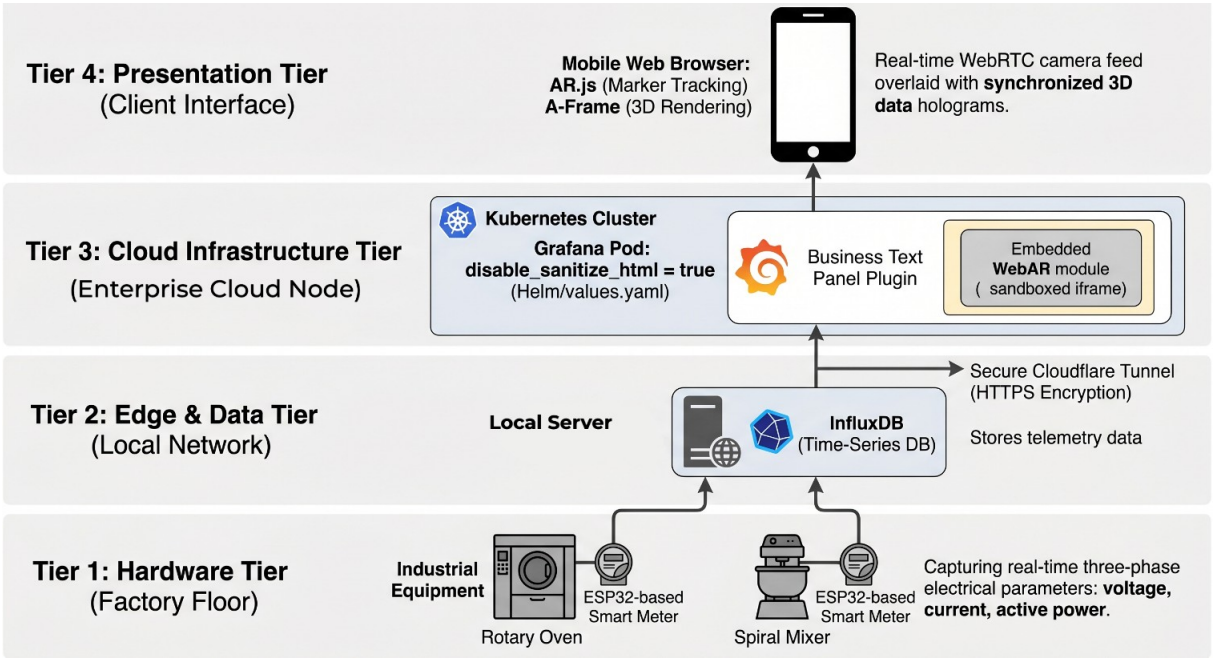


Fig. 1. Architecture of the WebAR industrial monitoring system and its secure data flow

At the foundational hardware tier, ESP32-based smart meters capture critical electrical parameters from industrial equipment [10]. This telemetry is transmitted to the edge and data tier, utilizing a local InfluxDB instance, which provides the necessary throughput and low-latency querying required for dynamic digital twins [3]. Because the factory server operates internally, a zero-trust Cloudflare Tunnel securely exposes the data to the cloud. This tunnel provides the mandatory HTTPS encryption required for WebRTC camera access without compromising internal network security [11]. The cloud infrastructure tier hosts the core visualization engine on a Kubernetes cluster, utilizing Grafana as the primary dashboarding platform [12]. To bypass Grafana's strict XSS security policies that actively block iframe executions, the deployment was modified at the Helm chart level by disabling the HTML sanitizer. The AR module is subsequently embedded using the Business Text plugin, safely rendering the sandboxed iframe containing the AR tracking logic while maintaining overall pod integrity.

Finally, the presentation tier enables zero-install access via standard mobile web browsers, circumventing platform-specific ecosystem restrictions [6]. The WebAR client uses AR.js for resilient marker tracking and A-Frame for rendering 3D holograms. It asynchronously fetches sensor data from InfluxDB via Flux queries and dynamically overlays the metrics onto the physical equipment in the camera's viewport, ensuring near real-time operational awareness [13].

3. Implementation and algorithms

The successful integration of WebAR into the monitoring dashboard required addressing two primary technical challenges: real-time anomaly detection at the edge and spatial stabilization of the holographic user interface.

3.1. Real-time data acquisition and anomaly detection

The WebAR client operates asynchronously, fetching telemetry data from the InfluxDB instance every 5 seconds. To minimize payload size and optimize mobile network performance, the system utilizes the Flux query language to request only the latest scalar values for targeted device IDs, filtering specifically for active power, voltage, and multi-phase currents [14]. Beyond simple data visualization, the AR module performs on-device edge computing to detect critical mechanical anomalies, specifically phase imbalance in three-phase industrial motors. A significant phase imbalance indicates issues such as degraded contacts, uneven load distribution, or partial phase loss, which can lead to catastrophic overheating and stator winding

damage. To ensure scalability across equipment of varying power capacities, the algorithm calculates the relative current unbalance percentage rather than relying on static thresholds. The system evaluates the real-time currents (I_a, I_b, I_c) to compute the average current I_{avg} and the maximum deviation from this average. The condition for triggering a critical phase imbalance alert is defined mathematically as follows:

$$I_{avg} = \frac{I_a + I_b + I_c}{3} \tag{1}$$

$$Imbalance\ (\%) = \frac{max(|I_a - I_{avg}|,\ |I_b - I_{avg}|,\ |I_c - I_{avg}|)}{I_{avg}} \times 100\% \tag{2}$$

$$Alert = \begin{cases} True, & if \max(I_a,\ I_b,\ I_c) > 1.0\ A\ and Imbalance > 15\% \\ False, & otherwise \end{cases} \tag{3}$$

If this condition is met, the AR module dynamically updates the Document Object Model (DOM), altering the holographic frame's emissive color to red and displaying a critical “PHASE IMBALANCE” alert directly over the physical machine. A secondary algorithm queries 26-hour historical minimums to detect equipment running non-stop, triggering operational warnings to prevent accelerated wear.

3.2. Holographic UI stabilization

A significant limitation of marker-based AR tracking (e.g., AR.js) is that the generated 3D coordinate system is rigidly bound to the physical marker's orientation. If a physical marker is placed on a machine at an angle or upside down, the projected data panels become misaligned and difficult for the engineer to read.

To solve this, a custom A-Frame component named always-upright was developed. This component applies a “billboard effect,” ensuring the data panel always faces the user's camera regardless of the physical marker's pitch, yaw, or roll. This requires calculating the relative orientation using quaternion mathematics, avoiding the “gimbal lock” problem inherent to standard Euler angles [15]. Let q_c represent the world quaternion of the mobile device's camera, and q_m represent the world quaternion of the physical marker. To ensure the holographic plane faces the camera, its local quaternion q_p must negate the marker's rotation and apply the camera's rotation. The required transformation is calculated by multiplying the inverse of the marker's quaternion by the camera's quaternion:

$$q_p = q_m^{-1} \otimes q_c \tag{4}$$

This mathematical transformation is executed continuously on every render tick within the WebGL pipeline. By dynamically updating the local rotation of the holographic entity based on this formula, the system achieves a highly stable UI. This significantly improves the user experience, allowing engineers to scan equipment from various angles and distances without losing the readability of the telemetry data.

4. Results and discussion

The proposed WebAR-integrated Grafana system was deployed and evaluated on a live bakery factory floor, monitoring 16 critical pieces of equipment over a continuous 30-day period. The evaluation focused on system performance, architectural reliability, and operational impact.

4.1. System performance and latency

Real-time monitoring requires minimal latency between data acquisition and visual rendering. As demonstrated in earlier evaluations, InfluxDB is highly capable of managing high-frequency telemetry, natively processing up to 8,000 metrics per second [3]. In our deployment, the end-to-end latency – from the ESP32 smart meter transmitting data, through the Cloudflare Tunnel, into the Kubernetes-hosted Grafana instance, and finally rendered as a WebAR hologram on the mobile client – averaged under 800 milliseconds over a 4G/LTE cellular network. This near-instantaneous data delivery is more than sufficient for industrial

anomaly detection, ensuring that critical events, such as a phase imbalance, are visualized on the engineer's device almost immediately upon occurrence.

To illustrate the underlying data pipeline that enables this rapid response, Fig. 2 presents the raw three-phase current telemetry (I_a, I_b, I_c) captured natively within the InfluxDB engine during such an event. The time-series graph explicitly demonstrates the sudden electrical imbalance – where one phase drops near zero while the others spike. Because the system processes these high-frequency scalar values with sub-second latency, the AR module can evaluate the mathematical conditions in real time and trigger the corresponding holographic alerts without noticeable delay.

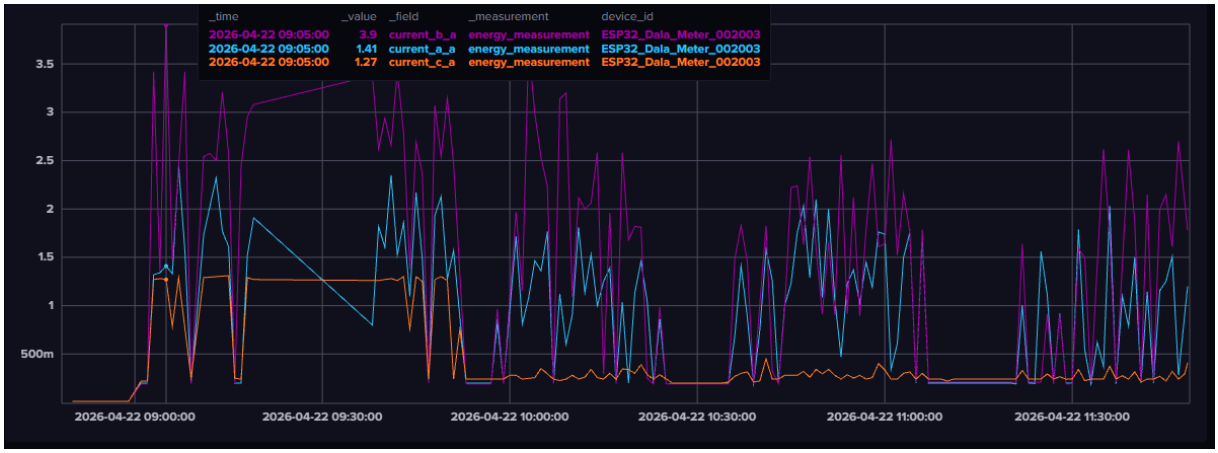


Fig. 2. Real-time three-phase current telemetry captured via the InfluxDB engine, demonstrating the sudden electrical imbalance detected by the system prior to the AR alert generation

4.2. Architectural reliability

The integration of a custom WebAR module into an enterprise Kubernetes cluster inherently introduces security and stability concerns. However, the architectural decision to disable the HTML sanitizer specifically for the Business Text plugin proved to be both stable and secure. By isolating the AR execution logic within a sandboxed iframe, the overarching Grafana pod maintained its structural integrity and was unaffected by the heavy WebGL rendering occurring on the client side. Furthermore, the Cloudflare Tunnel maintained a 99.9% uptime, successfully bridging the localized, unencrypted factory network with the strict HTTPS requirements of mobile WebRTC without exposing internal IP addresses to the public internet.

4.3. Operational impact and user experience

The most significant outcome of the deployment was the reduction in cognitive load for maintenance personnel. By eliminating the need to manually correlate physical machines with remote 2D/3D dashboards [4], the zero-install WebAR interface streamlined operations. The quaternion-based “billboard effect” resolved marker misalignment, allowing engineers to scan equipment effectively from sharp angles without hologram skewing. Furthermore, the localized edge-computing algorithms successfully identified simulated phase imbalance events, proving the system's viability as a proactive diagnostic tool. To quantify the operational impact, a preliminary user evaluation was conducted with maintenance engineers over the 30-day deployment. Participants reported a noticeable reduction in average fault localization time – decreasing from approximately 4.5 minutes using traditional remote dashboards to 2.5 minutes using the WebAR interface. In a post-deployment feedback session, users highlighted the system's ease of use, confirming that the zero-install approach significantly reduced the barrier to utilizing spatial diagnostics on the factory floor. The system overlays dynamic, color-coded holographic frames to indicate operational states: real-time telemetry displays a green frame under normal conditions (Fig. 3a), while detected anomalies, such as a critical phase imbalance, trigger an immediate red spatial alert (Fig. 3b). In addition to live telemetry, each hologram exposes a static equipment passport (manufacturer, model, specification, rated power and serial number) retrieved from the asset registry, so that the engineer can verify the machine identity

and rated parameters without leaving the AR view (Fig. 3c). The secondary duty-cycle algorithm described in Section 3.1 is illustrated in Fig. 3d: an ice maker whose current has not dropped below the idle threshold for more than 26 hours is flagged with a “NON-STOP > 26H” warning, and the frame colour switches from green to amber to distinguish this operational warning from the critical red alert.

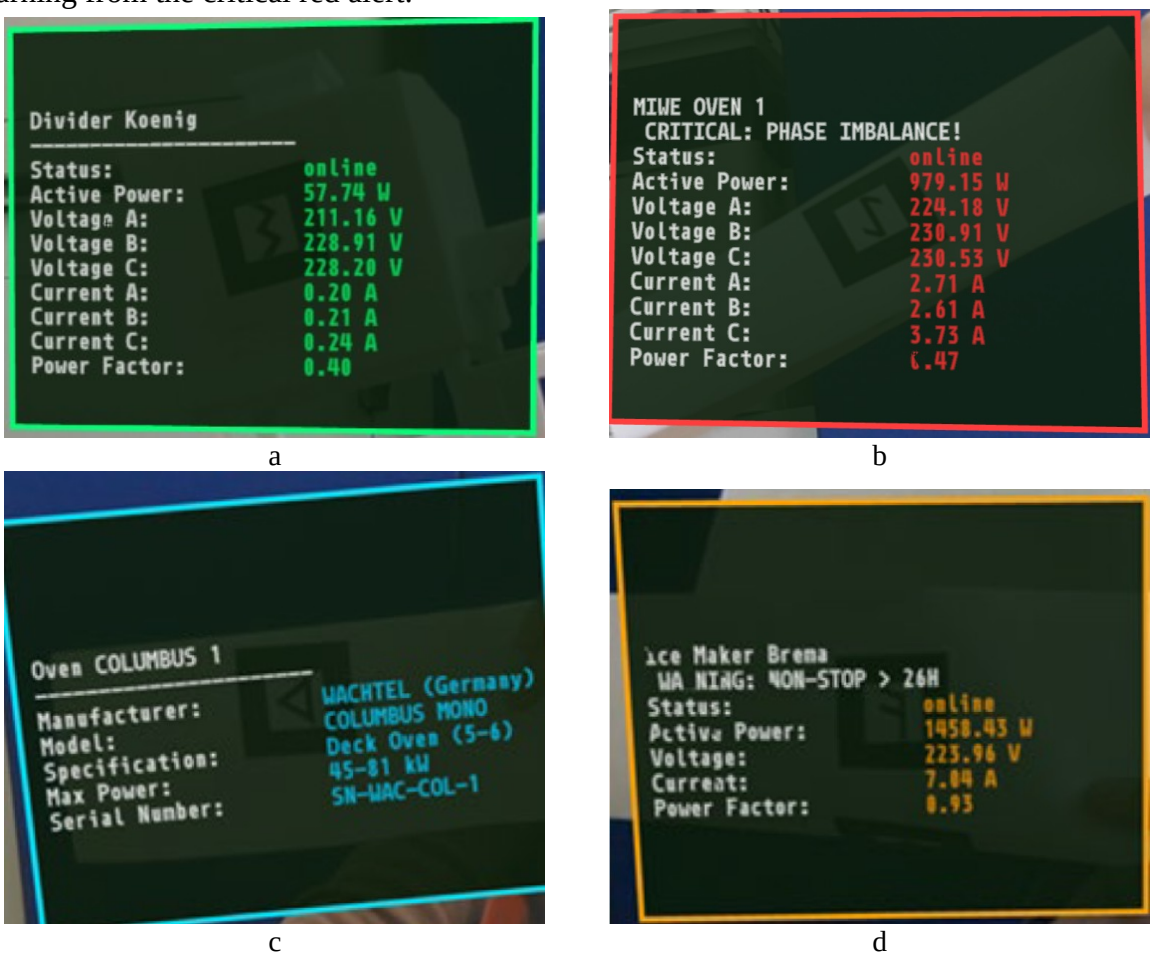


Fig. 3. Holographic panels of the mobile WebAR interface on the factory floor: (a) standard telemetry of a machine in normal operation (green frame); (b) critical alert triggered by the phase imbalance detection algorithm (red frame); (c) equipment passport of a deck oven; (d) operational warning for an ice maker running non-stop for more than 26 hours (amber frame)

Compared to traditional 2D SCADA and standard Grafana dashboards, which force engineers to context-switch between abstract remote metrics and physical machinery, the proposed system overlays telemetry directly onto the asset. Unlike native industrial AR solutions that necessitate expensive hardware or strict Mobile Device Management (MDM) policies for installation, our WebAR approach operates entirely within standard mobile browsers. This eliminates deployment friction and significantly lowers the cost of adoption, providing a highly accessible alternative to heavy spatial computing platforms. Despite its significant advantages, the marker-based WebAR approach presents specific limitations in harsh industrial environments. The tracking stability of the AR.js engine relies heavily on visual contrast. Consequently, poor ambient lighting on the factory floor or the accumulation of dust, flour, and grease on the physical markers can degrade hologram registration. Additionally, the system requires the engineer to maintain a direct line of sight with the marker, which can be restrictive in cramped machinery spaces.

Regarding system scalability, while the current deployment was evaluated on 16 assets, the underlying cloud-native architecture (InfluxDB and Kubernetes) is theoretically capable of processing high-frequency telemetry from hundreds of machines simultaneously, based on established database benchmarks [3]. However, the primary scalability bottleneck resides on the

client side. Rendering multiple dynamic 3D holograms simultaneously via WebGL can strain the thermal and processing limits of standard mobile devices.

5. Conclusion

This paper presented a detailed architectural framework for integrating Web-based Augmented Reality (WebAR) directly into strict enterprise monitoring platforms like Grafana. By successfully navigating Kubernetes security protocols, solving WebRTC network limitations via zero-trust tunnels, and developing custom quaternion mathematics for holographic UI stabilization, a reliable, zero-install AR dashboard was realized.

The deployment on a live factory floor demonstrated that combining the analytical power of InfluxDB with the spatial context of AR.js significantly enhances real-time industrial monitoring. The system bridges the gap between raw IIoT telemetry and physical machinery, enabling rapid, on-site anomaly detection such as phase imbalance. Future work will explore markerless tracking techniques utilizing WebXR APIs and the integration of machine learning models directly into the AR overlay for predictive maintenance visualizations.

Acknowledgements

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. BR24992975 “Development of a Digital Twin for the Food Industry Enterprise Using Artificial Intelligence and IIoT Technologies”).

References

1. Leng, J. Digital twins-based smart manufacturing system design in Industry 4.0: A review / J. Leng, D. Wang, W. Shen, X. Li, Q. Liu, X. Chen // Journal of Manufacturing Systems. – 2021. – Vol. 60. – P. 119–137. – DOI: 10.1016/j.jmsy.2021.05.011.

2. Jamil, M.N. Enabling Industrial Internet of Things by leveraging distributed edge-to-cloud computing: Challenges and opportunities / M.N. Jamil, O. Schelén, A.A. Monrat, K. Andersson // IEEE Access. – 2024. – Vol. 12. – P. 127294–127308. – DOI: 10.1109/ACCESS.2024.3454812.

3. Amirkhanov, B. Digital twins and IIoT: comparison of Prometheus and InfluxDB / B. Amirkhanov, T. Ishmurzin, M. Kunelbayev, G. Amirkhanova, A. Aidynuly, G. Tyulepberdinova // Bulletin of Electrical Engineering and Informatics. – 2025. – Vol. 14, N. 5. – DOI: 10.11591/eei.v14i5.9687.

4. Amirkhanov, B. Creating 3D models of production equipment and infrastructure using Blender / B. Amirkhanov, T. Nurgazy, G. Amirkhanova, M. Kunelbayev, G. Tyulepberdinova // International Journal of Innovative Research and Scientific Studies. – 2025. – Vol. 8, N. 1. – P. 1572–1588. – DOI: 10.53894/ijirss.v8i1.4704.

5. Egger, J. Augmented reality in support of intelligent manufacturing – A systematic literature review / J. Egger, T. Masood // Computers & Industrial Engineering. – 2020. – Vol. 140. – Art. 106195. – DOI: 10.1016/j.cie.2019.106195.

6. Qiao, X. Web AR: A promising future for mobile augmented reality – State of the art, challenges, and insights / X. Qiao, P. Ren, S. Dustdar, L. Liu, H. Ma, J. Chen // Proceedings of the IEEE. – 2019. – Vol. 107, N. 4. – P. 651–666. – DOI: 10.1109/JPROC.2019.2895105.

7. Vaño, R. Cloud-native workload orchestration at the edge: A deployment review and future directions / R. Vaño, I. Lacalle, P. Sowiński, R. S-Julián, C.E. Palau // Sensors. – 2023. – Vol. 23, N. 4. – Art. 2215. – DOI: 10.3390/s23042215.

8. Cauz, M. Text readability in augmented reality: a multivocal literature review / M. Cauz, A. Clarinval, B. Dumas // Virtual Reality. – 2024. – Vol. 28, N. 1. – Art. 59. – DOI: 10.1007/s10055-024-00949-6.

9. Boyes, H. The industrial internet of things (IIoT): An analysis framework / H. Boyes, B. Hallaq, J. Cunningham, T. Watson // Computers in Industry. – 2018. – Vol. 101. – P. 1–12. – DOI: 10.1016/j.compind.2018.04.015.

10. Saleem, M.U. Design, implementation, and deployment of an IoT based smart energy management system / M.U. Saleem, M.R. Usman, M. Shakir // IEEE Access. – 2021. – Vol. 9. – P. 59649–59664. – DOI: 10.1109/ACCESS.2021.3070960.

11. Syed, N.F. Denial of service attack detection through machine learning for the IoT / N.F. Syed, Z. Baig, A. Ibrahim, C. Valli // Journal of Information and Telecommunication. – 2020. – Vol. 4, N. 4. – P. 482–503. – DOI: 10.1080/24751839.2020.1767484.

12. Calderón, D. Implementation and experimental application of industrial IoT architecture using automation and IoT hardware/software / D. Calderón, F.J. Folgado, I. González, A.J. Calderón // Sensors. – 2024. – Vol. 24, N. 24. – Art. 8074. – DOI: 10.3390/s24248074.

13. Mota, J.M. Augmented reality mobile app development for all / J.M. Mota, I. Ruiz-Rube, J.M. Dodero, I. Arnedillo-Sánchez // Computers & Electrical Engineering. – 2018. – Vol. 65. – P. 250–260. – DOI: 10.1016/j.compeleceng.2017.08.025.

14. Grzesik, P. Comparative analysis of time series databases in the context of edge computing for low power sensor networks / P. Grzesik, D. Mrozek // Computational Science – ICCS 2020. Lecture Notes in Computer Science. – Cham: Springer, 2020. – Vol. 12139. – P. 371–383. – DOI: 10.1007/978-3-030-50426-7_28.

15. Vince, J. Quaternions for computer graphics / J. Vince. – 2nd ed. – London: Springer, 2021. – 181 p. – DOI: 10.1007/978-1-4471-7509-4.

Information about the authors

Nurgazy Tomiris – PhD student in Intelligent Control Systems, researcher, Laboratory of Digital Twins, Department of Artificial Intelligence and Big Data, Al-Farabi Kazakh National University, Almaty, Kazakhstan; e-mail: nurgazytomiris@gmail.com; ORCID: 0009-0001-6650-5517

Amirkhanova Gulshat – PhD in Computer Science, Senior Lecturer, Department of Artificial Intelligence and Big Data, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan; e-mail: gulshat.aa@gmail.com; ORCID: 0000-0003-3933-5476

Amirkhanov Bauyrzhan – PhD student, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan; Senior Lecturer, Turan University, Almaty, Kazakhstan; e-mail: amirkhanov.b@gmail.com; ORCID: 0000-0002-4915-0347

Rustemov Almas – Master of Natural Sciences, Lecturer, Department of Electronics and Astrophysics, Al-Farabi Kazakh National University, Almaty, Kazakhstan; ORCID: 0009-0009-6970-0746

Liu Huanyu – researcher, Peking University, Beijing, China; e-mail: huanyuliu@stu.pku.edu.cn; ORCID: 0009-0001-2183-9716

FORECASTING CROP YIELDS USING XGBOOST AND BAYESIAN HYPERPARAMETER OPTIMIZATION

A. Zhussip[1], G. Ziyatbekova[2,3], T. Sarsembayeva[2], A. Akhmetova[2], A. Tengaeva[1], N. Tasbolatuly[4]

[1]Kazakh National Agrarian Research University, Almaty, Kazakhstan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[3]Almaty Technological University, Almaty, Kazakhstan
[4]Astana International University, Astana, Kazakhstan
E-mail: alisher123841@bk.ru

Abstract. *Food security requires increasing the efficiency of agricultural production. In the context of climate change and soil degradation, accurate crop yield forecasting is becoming increasingly important. Machine learning methods make it possible to analyze climatic, soil, and agronomic factors affecting crop yields. The study was conducted using the Agriculture Crop Yield dataset, which includes climatic, soil, and agronomic features. Categorical features were transformed using One-Hot Encoding, while numerical features were standardized using StandardScaler. The data were divided into training and test sets at a ratio of 80:20. The study compared a baseline XGBoost model with a model using Bayesian hyperparameter optimization implemented with Optuna and the TPE sampler over 50 trials. Model performance was evaluated using MAE, RMSE, and R^2 metrics. The baseline XGBoost model achieved MAE = 0.5014 t/ha, RMSE = 0.6267 t/ha, and R^2 = 0.8638. After Bayesian optimization, the values improved to MAE = 0.3998 t/ha, RMSE = 0.5011 t/ha, and R^2 = 0.9129. Thus, MAE decreased by 20.26%, RMSE by 20.05%, while R^2 increased by 5.69%. The obtained results demonstrate improved forecasting performance when using an optimized XGBoost configuration. Bayesian hyperparameter optimization improved the performance of the XGBoost model on the considered dataset. The optimized model demonstrated better results across all three evaluation metrics. The results confirm the research hypothesis that automatic hyperparameter tuning provides advantages over a fixed configuration. However, the conclusions are limited to the dataset used and the conducted experiment. In future research, it would be appropriate to use cross-validation and test the approach on additional datasets.*

Keywords: *machine learning, crop yield forecasting, XGBoost, gradient boosting, Bayesian optimization, hyperparameter optimization, Optuna, TPE, precision agriculture.*

Introduction. Global food security remains one of the key challenges of the 21st century. According to the UN Food and Agriculture Organization (FAO), the world population will exceed 9.1 billion by 2050, requiring at least a 70% increase in food production. Amid climate change, soil degradation and shrinking arable land, traditional farming methods are proving insufficient to meet this challenge [4].

Against the backdrop of rapid development in information technology, machine learning is being increasingly adopted in agriculture. The ability to analyze large volumes of data climatic indicators, soil characteristics, agronomic parameters and build accurate yield forecasts opens new opportunities for optimizing agricultural production.

Of particular interest is an analysis of existing methodological approaches to yield forecasting. Section 2 discusses, as an example, a methodology applied in Germany for forecasting winter-wheat yield, based on a combination of deep neural networks, XGBoost and model-interpretability methods (SHAP). This material serves as context for existing research (related work) and is not part of the experimental portion of the present study.

The aim of this study is to develop and experimentally compare two yield-forecasting models based on XGBoost: (1) a baseline model with pre-set hyperparameters, and (2) a model whose hyperparameters are automatically selected via Bayesian Hyperparameter Optimization.

The main hypothesis is that automatic Bayesian hyperparameter selection yields a more accurate, better-generalizing model than a fixed, manually specified configuration. Bayesian Optimization here is applied exclusively to tuning XGBoost's hyperparameters and is not itself a forecasting model [1].

Methods. Before training, the source dataset was checked for missing values (df.isna().sum()), full duplicate rows (df.duplicated().sum()), and whether numerical features fell within physically plausible ranges. For the numerical features Rainfall_mm, Temperature_Celsius and Days_to_Harvest, standardization was applied (Formula 1): where x is the original value, μ the mean and σ the standard deviation, both computed on the training set.

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

The categorical features Region, Soil_Type, Crop and Weather_Condition were transformed via One-Hot Encoding (drop_first=True), removing redundancy and the dummy-variable trap. Boolean features Fertilizer_Used and Irrigation_Used were used as is. Data were then split 80/20 (random_state = 42) into training and test sets; the test set was reserved solely for final evaluation. For hyperparameter optimization, the training set was further split 80/20 into training/validation subsets (random_state = 42) (Table 1).

Table 1. Baseline XGBoost hyperparameters

№	**Hyperparameter**	**Value**
1	n_estimators	30
2	learning_rate	0.05
3	max_depth	5
4	random_state	42

XGBoost is an ensemble algorithm that sequentially builds decision trees; each new tree is trained on the errors of the previous ensemble, and the final prediction sums the contributions of individual trees. The optimization objective includes both a loss function and a regularization term limiting model complexity. XGBoost was chosen for its effectiveness with tabular data and its ability to model nonlinear relationships between agrometeorological and agronomic features [2, 5].

In Experiment 2, XGBoost hyperparameters were tuned using Optuna and the TPE (Tree-structured Parzen Estimator) algorithm. Unlike Grid Search (predefined combinations) or Random Search (independent configurations), TPE uses the results of previous trials to identify promising regions of the hyperparameter space. In each trial, XGBoost was trained on X_train_opt/y_train_opt and evaluated on X_valid/y_valid via MAE; the result guided Optuna's choice of the next configuration (Table 2).

Table 2. Search space for Bayesian hyperparameter optimization

№	**Hyperparameter**	**Search range**
1	n_estimators	50–500
2	learning_rate	0.01–0.3
3	max_depth	2–10
4	subsample	0.6–1.0
5	colsample_bytree	0.6–1.0
6	min_child_weight	1–10
7	gamma	0–5
8	reg_alpha	0–1
9	reg_lambda	0.1–10

Three regression metrics were used: MAE (Formula 2), RMSE (Formula 3) and R^2 (Formula 4). For MAE and RMSE, lower values indicate more accurate forecasts; RMSE is more sensitive to large errors due to squared deviations. R^2 reflects the share of target-variable variance explained by the model, with values closer to 1 indicating higher quality [3].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4}$$

Materials. The experimental part of the study is based on the tabular Agriculture Crop Yield dataset (Table 3), containing yield observations for various crops as a function of agrometeorological and agronomic factors. The target variable was crop yield (Yield_tons_per_hectare, t/ha).

Table 3. Dataset features

№	**Feature**	**Type**	**Description / values**
1	Region	categorical	North, East, South, West
2	Soil_Type	categorical	Clay, Sandy, Loam, Silt, Peaty, Chalky
3	Crop	categorical	Wheat, Rice, Maize, Barley, Soybean, Cotton
4	Rainfall_mm	numerical	precipitation, mm
5	Temperature_Celsius	numerical	temperature, °C
6	Fertilizer_Used	boolean	True / False
7	Irrigation_Used	boolean	True / False
8	Weather_Condition	categorical	Sunny, Rainy, Cloudy
9	Days_to_Harvest	numerical	days to harvest
10	Yield_tons_per_hectare	numerical	target variable, t/ha

The dataset thus combines categorical, numerical and boolean features, allowing the study of machine-learning algorithms for yield forecasting based on heterogeneous agronomic and weather factors.

As methodological context, existing German yield-forecasting studies were reviewed. The German approach uses deep neural networks and ensemble methods, particularly XGBoost, together with SHAP and TreeSHAP interpretation methods, drawing on meteorological, soil, phenological and statistical data brought to a common spatio-temporal scale. Meteorological data include temperature, solar radiation, precipitation, humidity and wind speed; soil data include particle-size composition, plant-available moisture and soil density; statistical data provide regional yield figures. This approach demonstrates the value of integrating heterogeneous factors but is computationally demanding and does not fully resolve model interpretability. The German experience is treated here as related work and methodological context; the present experiment uses a separate Agriculture Crop Yield dataset (Figure 1).

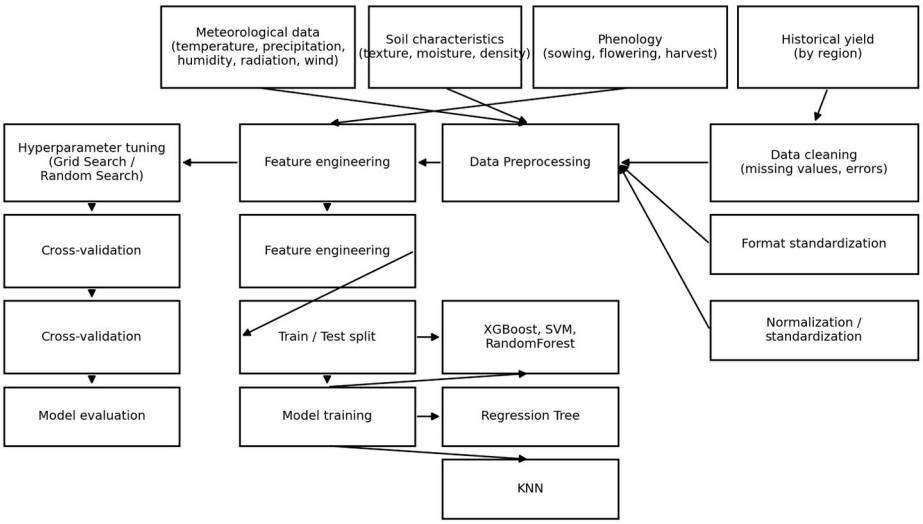


Figure 1. Generalized pipeline of the German yield-forecasting study (related work).

Results. The baseline XGBoost model (fixed hyperparameters) achieved, on the test set, MAE = 0.501390, RMSE = 0.626689 and R^2 = 0.863771. After Bayesian optimization with Optuna/TPE, the configuration in Table 4 was obtained.

Table 4. Optimized hyperparameter configuration

№	**Hyperparameter**	**Value**
1	n_estimators	337
2	learning_rate	0.0562
3	max_depth	5
4	subsample	0.7417
5	colsample_bytree	0.6461
6	min_child_weight	2
7	gamma	1.7961
8	reg_alpha	0.9193
9	reg_lambda	1.8655

The best validation MAE was 0.400157. The final model, trained on the full training set, achieved on the test set MAE = 0.399792, RMSE = 0.501055 and R^2 = 0.912917 (Table 5).

Table 5. Test-set results

№	**Model**	**MAE**	**RMSE**	**R^2**
1	XGBoost Baseline	0.501390	0.626689	0.863771
2	XGBoost + Bayesian Optimization	0.399792	0.501055	0.912917

Hyperparameter optimization reduced MAE by 20.26% and RMSE by 20.05%, and increased R^2 by 5.69%; the optimized model therefore showed higher yield-forecasting accuracy across all metrics. Over 50 trials, Optuna progressively explored the hyperparameter space, finding configurations with progressively lower validation MAE; the best value obtained was 0.400157.

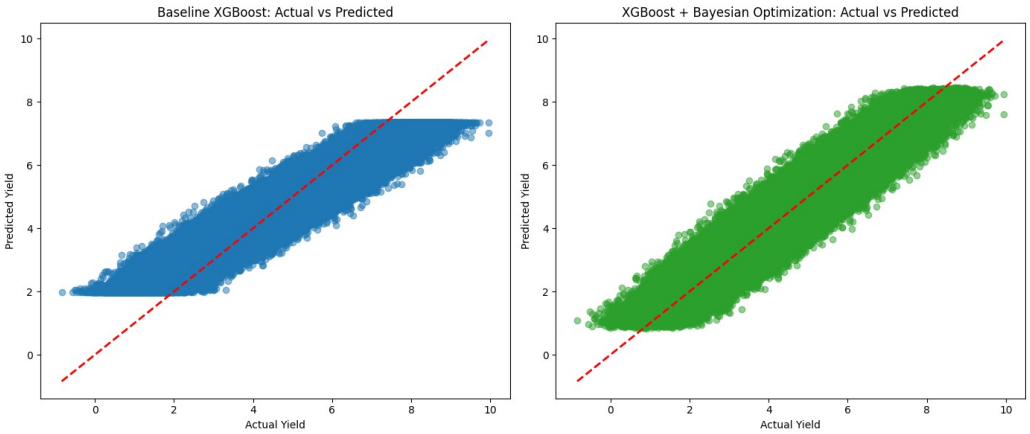


Figure 4. Convergence of Bayesian optimization over 50 trials.

To assess the influence of input factors on the yield forecast, feature importance was computed for the baseline and optimized XGBoost models. Comparing the top-10 features shows whether the contribution of individual agrometeorological and agronomic factors changed after hyperparameter optimization.

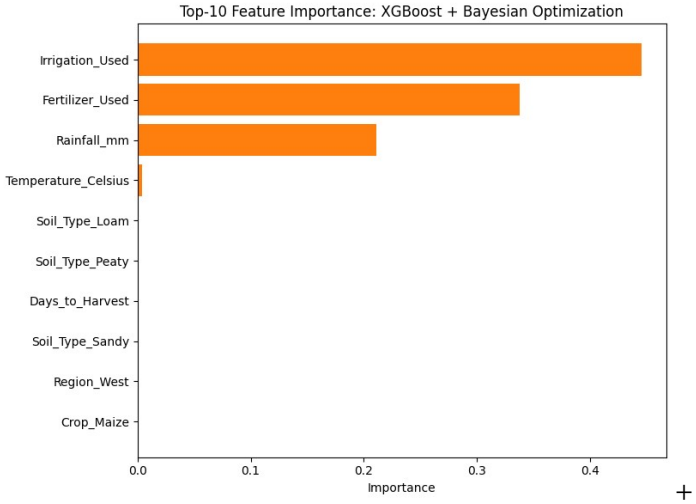


Figure 5. Top-10 most important features for the baseline and optimized XGBoost models.

Model quality was additionally assessed visually by comparing actual and predicted yield values; the closer the points lie to the ideal-prediction line $y = \hat{y}$, the higher the model's accuracy.

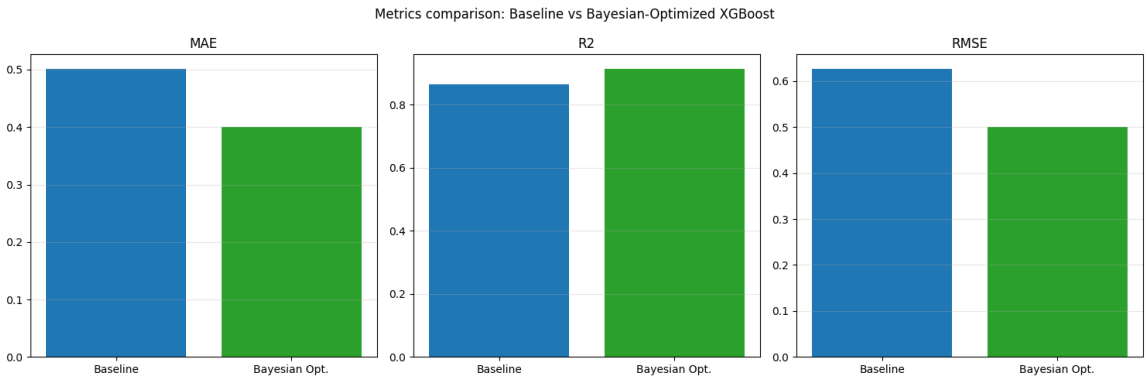


Figure 6. Comparison of actual and predicted yield values.

Discussion. The improvement after Bayesian optimization is not attributable to a single parameter but to the joint retuning of the entire hyperparameter set, affecting both ensemble complexity and regularization. Besides n_estimators, max_depth and learning_rate, subsample, colsample_bytree, min_child_weight, gamma, reg_alpha and reg_lambda were also optimized controlling the fraction of observations/features per tree, minimum leaf weight, the minimum gain threshold for a split, and L1/L2 regularization of tree weights.

Notably, the optimized configuration kept tree depth at the baseline level (max_depth = 5) and a learning_rate close to baseline (0.056 vs. 0.05), but substantially increased the number of trees (n_estimators = 337 vs. 30). This suggests that with a small learning rate and moderate depth, more boosting iterations are needed for acceptable quality, while overfitting risk from more trees is partly offset by the regularization found (gamma ≈ 1.80, reg_alpha ≈ 0.92, reg_lambda ≈ 1.87) and sampling parameters (subsample ≈ 0.74, colsample_bytree ≈ 0.65, min_child_weight = 2).

These observations are descriptive and do not establish a causal link between any single hyperparameter and the quality gain in isolation the effect results from the joint configuration found. Isolating individual contributions would require additional sensitivity analysis (e.g., an ablation study), outside this experiment's scope.

In this experiment, Bayesian optimization of XGBoost hyperparameters produced a consistent improvement across all three quality metrics on the test set, consistent with the general premise of Bayesian Optimization: sequentially using prior-trial results to target more promising regions of the hyperparameter space is more efficient than random or exhaustive search under a limited trial budget (here, 50 trials).

Conclusions. This study compared a baseline XGBoost model with an XGBoost model tuned via Bayesian optimization (Optuna, TPE sampler), evaluated on an independent test set. Hyperparameter optimization improved yield-forecasting accuracy: MAE decreased from 0.5014 to 0.3998 t/ha, RMSE from 0.6267 to 0.5011 t/ha, and R^2 increased from 0.8638 to 0.9129 the optimized model outperformed the baseline on every metric.

These results confirm the effectiveness of Bayesian optimization for tuning XGBoost on this dataset. The conclusions are, however, specific to this dataset and experimental setup. Future research could employ cross-validation, increase the number of trials, analyze feature importance in more depth, and test the method on additional yield datasets.

References

1. Smirnova, V. S., Shalamov, V. V., Efimova, V. A., & Filchenkov, A. A. (2020). Hyperparameter optimization based on combining a priori and a posteriori knowledge about the classification problem. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, *20*(6), 828–834. https://cyberleninka.ru/article/n/optimizatsiya-giperparametrov-na-osnove-obedineniya-apriornyh-i-aposteriornyh-znaniy-o-zadache-klassifikatsii/viewer

2. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

3. Kalichkin, V. K., Fedorov, D. S., Maksimovich, K. Yu., & Riksen, V. S. (2025). Machine learning for predicting crop rotation productivity. *Agroengineering, 27*(1), 41–52. https://doi.org/10.26897/2687-1149-2025-1-41-52

4. Food and Agriculture Organization of the United Nations. (2009). *Feeding the world in 2050*. FAO. https://www.fao.org/4/k6021e/k6021e.pdf

5. Srivastava, Amit & Safaei, Nima & Khaki, Saeed & Lopez, Gina & Zeng, Wenzhi & Ewert, Frank & Gaiser, Thomas & Rahimi, Jaber. (2021). Comparison of Machine Learning Methods for Predicting Winter Wheat Yield in Germany. 10.48550/arXiv.2105.01282.

https://www.researchgate.net/publication/351342435_Comparison_of_Machine_Learning_Methods_for_Predicting_Winter_Wheat_Yield_in_Germany

Дарынды оқушылармен жұмыс істеуде инновациялық педагогикалық технологияларды қолдану

Мадьярова Г.А. Орынбаева А.С.
aizhan.sansyzbaevna@mail.ru
Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

Аңдатпа. *Қазіргі білім беру жүйесінде, дарынды оқушылармен жұмыс істеуде инновациялық педагогикалық технологияларды қолдану қарастырылған. Мақсаты дарынды оқушыларды оқытуда инновациялық педагогикалық технологияларды қолданудың тиімділігін айқындау және сараптау. Зерттеу барысында әдебиеттерге шолу және сауалнама әдістері қолданылды. Нәтижесінде диагностикалық тест пен практикалық тапсырмалар тиімді әдістер ретінде анықталды. Қабілетті дамытуда жеке оқу траекториясы мен шығармашылық-жобалық жұмыстардың маңызы жоғары болды. Оқушы қабілетін анықтауда мұғалімнің рөлі басым екені анықталды. Интерактивті оқытуда Peer Instruction және ChatGPT сияқты заманауи технологияларды қолдану қажеттілігі анықталды. Зерттеу инновациялық технологияларды дұрыс пайдалану дарынды оқушылардың шығармашылық және сыни ойлау қабілеттерін дамытуға мүмкіндік беретінін көрсетті.*

Кіріспе. Білім жүйесінде дарынды немесе ынтасы жоғары оқушылардың қабілетін анықтап, оны дамытуда инновациялық технологияларды тиімді қолдану өзекті мәселелердің бірі болып отыр. Дарынды оқушылардың қызығушылығы мен мүмкіндігі әртүрлі болғандықтан, олармен жұмыс жасауда жеке әдіс қажет. Заманауи технологиялар оқушылардың шығармашылық, сыни ойлау және зерттеу қабілеттерін дамытуға мүмкіндік береді. Сондықтан мұғалімдер үшін тиімді әдістер мен цифрлық құралдарды дұрыс таңдау өте маңызды.

BİLSEM (Bilim ve Sanat Merkezi - Ғылым және өнер орталығы) орталықтарында оқитын 6–10 сыныптағы ерекше қабілетті оқушыларға жасанды интеллектті (AI) оқытуға арналған бастапқы деңгейдегі оқу бағдарламасының құрылымын әзірленген. Дизайнға негізделген зерттеу әдісі қолданылып, алдымен 41 университеттің 54 AI және машиналық оқыту оқу бағдарламасына контент-талдау жасалған, кейін AI саласындағы 20 сарапшының пікірі алынған. Сарапшылардың пікірлері мазмұнның валидтілік коэффициенті және иерархиялық кластерлік талдау арқылы өңделген. Нәтижесінде оқушыларға оқытуға ұсынылатын негізгі тақырыптар анықталып, 6 оқу саласы мен 22 ішкі оқу саласынан тұратын AI оқу бағдарламасының негізі жасалған. Тақырыптарға AI-ға кіріспе, ықтималдық, бақыланатын және бақыланбайтын оқыту, шешім ағаштары, кластерлеу және Naive Bayes жатады. Авторлар дарынды оқушыларға AI-ды тек дайын құралдар арқылы емес, оның математикалық, статистикалық, алгоритмдік және этикалық негіздерімен байланыстыра оқыту қажет екенін анықтаған. Бұл бағдарлама мұғалімдерге бағыт-бағдар береді, алайда оның оқушыларға нақты әсерін анықтау үшін кейін практикалық түрде сынақтан өткізу қажет екенін анықтады [1].

Gifted and talented (дарынды және талантты) оқушылардың компьютерлік бағдарламалауға өзіндік тиімділігі (computer programming self-efficacy) мен компьютерлік ойлау дағдылары (computational thinking) арасындағы байланысын зерттеген. Түркиядағы Балыкесир қаласының Science and Art Center орталығында білім алатын 106 орта мектеп оқушысы (46 қыз, 60 ұл) қатысты. Зерттеуде корреляциялық зерттеу әдісі қолданылып, деректер «Computational Thinking Skills Scale» және «Computer Programming Self-Efficacy Scale» арқылы жиналды; нәтижелер Пирсон корреляциясы және қарапайым сызықтық регрессия әдістерімен талданған. Бағдарламалауға өзіндік тиімділік пен компьютерлік ойлау арасында оң және орташа деңгейдегі байланыс анықталған ($r = .561$, $p < .01$). Бағдарламалауға өзіндік тиімділік оқушылардың компьютерлік ойлау дағдыларының 31,5%-ын түсіндіретін маңызды

болжамдаушы фактор екенін көрсеткен. Оқушының бағдарламалауды орындауға деген сенімі жоғары болған сайын, оның компьютерлік ойлау дағдылары да жоғары болуға бейім екенін көруге болады. Авторлар дарынды оқушыларды бағдарламалауға оқыту барысында олардың тек техникалық білімін емес, өзіндік сенімділігі мен мотивациялық-аффективтік дағдыларын да дамыту қажет деп санаған [2].

Kirmit Ş., Dönmez İ., & Çataltaş H. E. (2018). жоғары қабілетті (дарынды) орта мектеп оқушыларының компьютерлік ойлау дағдыларының жынысқа байланысты айырмашылығын зерттелген. Сипаттамалық және жалпы/өзара байланыстық сауалнама әдісімен жүргізілген. 2017–2018 оқу жылында Анкара қаласындағы Білім және өнер орталығында BİLSEM-түрік тіліндегі Bilim ve Sanat Merkezi (Ғылым және өнер орталығы) оқитын 59 дарынды оқушы: 26 қыз және 33 ұл қатысты. Оқушылардың компьютерлік ойлау дағдыларын анықтау үшін 22 сұрақтан тұратын, 5 компонентті Computational Thinking Skills Scale қолданылып, деректер SPSS 20 бағдарламасында өңделген. Деректер қалыпты таралмағандықтан, жыныстар арасындағы айырмашылықты анықтау үшін Mann–Whitney U тесті пайдаланылған. Нәтижесінде ұл оқушылардың шығармашылық ойлау, алгоритмдік ойлау және сыни ойлау көрсеткіштері қыздарға қарағанда статистикалық тұрғыдан жоғары екені анықталды ($p < .05$). Ал ынтымақтастықта оқу және проблеманы шешу дағдыларында жыныс бойынша маңызды айырмашылық байқалмаған. Оқушылардың ең жоғары көрсеткіші шығармашылық ойлауда (3,91), ал ең төмен көрсеткіші проблеманы шешуде (2,13) байқалды деп санаған. Авторлар дарынды оқушылардың, әсіресе проблеманы шешу дағдыларын дамытуға бағытталған білім беру жұмыстарын күшейту қажет деген қорытындыға келген [3].

Технологиялық құралдарды, соның ішінде жасанды интеллектіні, дұрыс пайдаланбаудың дарынды оқушылардың XXI ғасыр дағдыларына кері әсерін зерттеген. Технологиялық практикалардың дарынды оқушылардың мотивациясына, шығармашылығына, сыни ойлауына және проблеманы шешу қабілеттеріне ықтимал кері әсерлерін талдау. Зерттеу нысаны ретінде дарынды оқушыларға арналған 2010–2025 жылдары жарияланған ғылыми зерттеулер қарастырылған. Әдіс ретінде жүйелі әдеби шолу (systematic review) қолданылып, жұмыс PRISMA ұсынымдарына сәйкес жүргізілген. Бастапқы зерттеулердің ішінен талаптарға сәйкес келген 40 мақала соңғы талдауға енгізіліп, олардың зерттеу сапасы мен бұрмалану қаупі Cochrane RoB 2 және ROBINS-I құралдары арқылы бағаланған. Жаңа технологиялардың XXI ғасыр дағдыларын дамытуға үлкен мүмкіндігі бар екені анықталғанымен, оларды дұрыс ұйымдастырмай қолдану дарынды оқушылардың терең оқуға деген мотивациясын және шығармашылық белсенділігін төмендетуі, оқушылардың технологияны пассивті түрде қолдануына әкелуі мүмкін екені көрсетілген. Технологияны білім беруде қолдану кезінде оқушының сыни ойлауын, дербестігін, шығармашылығын және белсенді оқуын сақтауға ерекше көңіл бөлу қажет екенін дәлелдеген [4].

Дарынды оқушыларды анықтауда дәстүрлі IQ тестілерімен шектелмей, шығармашылық, аналитикалық, практикалық және даналық қабілеттерді кешенді бағалауды ұсынған. Автор бұл тәсілді Rainbow, Kaleidoscope және Aurora жобалары арқылы университетке түсушілер мен 10–12 жастағы оқушыларға қолданған. Нәтижелері кешенді бағалау оқушылардың академиялық жетістігін тиімдірек болжауға, дарындылықты кеңірек анықтауға және этникалық топтар арасындағы айырмашылықтарды азайтуға мүмкіндігін көрсетті. Sternberg R. J. дарынды оқушыларды анықтауда бір ғана интеллект көрсеткішіне емес, олардың әртүрлі қабілеттерінің өзара үйлесіміне сүйенуді зерттеу қажетілігн көрсеткен [5].

Дарынды оқушыны тек жоғары IQ немесе оқу жетістігі арқылы анықтау жеткіліксіз. Дарындылық баланың табиғи қабілетімен қатар, мотивация, мақсатты жаттығу, қолайлы білім беру ортасы, мұғалімнің қолдауы және берілетін мүмкіндіктер арқылы дамиды. Авторлар дарындылықты тұрақты қасиет емес, қабілеттің уақыт өте келе талант пен жоғары жетістікке айналу процесі ретінде қарастырды. Дарынды

оқушыларды анықтауда бірнеше бағалау әдісін қолданып, олардың қабілетін дамытуға бағытталған жеделдетілген және тереңдетілген білім беру қажеттілігін дәлелдеді [6].

Зерттеуінде дарынды 7-сынып оқушыларына Python бағдарламалау арқылы деректер ғылымын оқыту мүмкіндігі қарастырылған. Зерттеуге 11–12 жастағы алты дарынды оқушы қатысып, зерттеу барысында іс-әрекеттік зерттеу әдісі, мұғаліммен сұхбат, сабақтағы бақылау және оқушылардың жұмыстарына талдау қолданылған. Жеті апта бойы оқушылар статистикалық ұғымдарды Python, NumPy, Pandas және Matplotlib құралдары арқылы меңгеріп, нақты ойын деректерін талдаған. Нәтижесінде оқушылардың статистикалық ойлау, деректерді талдау және визуализациялау дағдылары дамып, нақты деректермен жұмыс жасау олардың қызығушылығы мен мотивациясын арттырғаны анықтаған. Авторлар деректер ғылымын бағдарламалаумен біріктіру дарынды оқушылардың деректер сауаттылығын дамытуда тиімділігін көрсетті [7].

Австралиядағы дарынды оқушыларға арналған АКТ саясатын зерттеп, ұлттық және штаттық деңгейдегі құжаттарды талдаған. Зерттеуде құжаттық талдау және әдебиеттерге шолу әдістері қолданылып, 6 саясаттық құжат пен 14 ғылыми мақала қарастырылған. Нәтижесінде дарынды оқушыларға арналған білім беру саясаты болғанымен, АКТ-ны қолдануға қатысты нақты әрі бірізді саясаттың жеткіліксіз екені анықталған. Әсіресе ауылдық және шалғай аймақтардағы оқушылардың АКТ-ға қолжетімділігінде теңсіздік байқалған. Авторлар дарынды оқушылардың білім сапасын арттыру үшін АКТ саясатын жетілдіру, мұғалімдердің цифрлық құзыреттілігін дамыту және білім алуға тең қолжетімділікті қамтамасыз ету қажет деп көрсеткен [8].

Дарынды оқушылардың жаратылыстану ғылымдарындағы «өсу ойлау жүйесі» (growth mindset), яғни интеллект пен қабілетті еңбектену және оқу арқылы дамытуға болады деген сенімі зерттелген. Зерттеуге 7-сыныптың 380 оқушысы қатысқан, оның ішінде 80-і дарынды және қабілетті оқушы ретінде анықталған. Квази-эксперименттік әдіс қолданылып, сауалнама және алты аптаға созылған Brainology компьютерлік интервенциясы жүргізілген; 32 дарынды оқушы интервенция тобына, 48 оқушы бақылау тобына бөлінген. Дарынды оқушылардың бастапқыда қарапайым оқушыларға қарағанда ғылымдағы интеллекттің өзгермелі екеніне сенімі жоғары болғанымен, олардың арасында да тұрақты (fixed) ойлау жүйесі бар оқушылар кездескен. Brainology бағдарламасына қатысқан дарынды оқушылардың өсу ойлау жүйесі статистикалық тұрғыдан едәуір жақсарған және бұл өзгеріс кейінгі бақылауда сақталған. Интервенцияның нәтижесіне мұғалімнің әсері анықталмаған. Зерттеу дарынды оқушылардың әлеуетін толық ашуда өсу ойлау жүйесін дамытуға бағытталған қысқа мерзімді цифрлық бағдарламалардың тиімді екенін көрсеткен [9].

Дарынды оқушыларды анықтау мен дамыту жағдайында информатика мұғалімдерінің кәсіби трансформациясы мәселесен қарастырған. Зерттеуде жүйелі контент-талдау, салыстырмалы педагогикалық талдау, педагогикалық диагностика, сауалнама, сұхбат, бақылау, сандық және сапалық талдау және модельдеу әдістері қолданылған. Нәтижесінде дарындылықты диагностикалау, оқушыларды қолдау және мұғалімдердің цифрлық құзыреттілігін дамыту бағыттары анықталып, осы бағыттарды біріктіретін Smart-Talent Core моделі ұсынылған. Мұғалімнің цифрлық және кәсіби құзыреттілігін арттыру дарынды оқушылардың интеллектуалдық, шығармашылық және жеке әлеуетін дамытуға мүмкіндік беретіні көрсетілген [10].

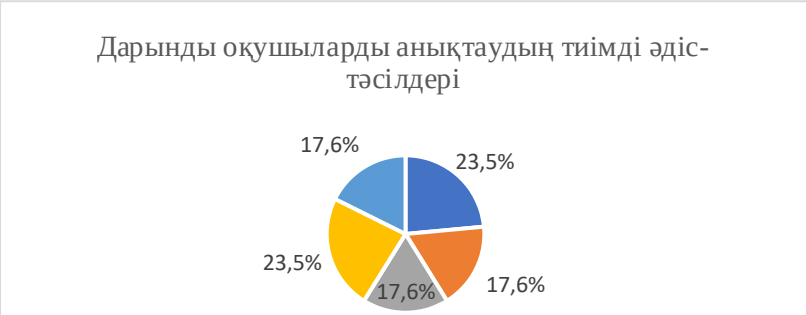
Зерттеу әдістері мен нәтижелері

Зерттеуде екі негізгі әдіс қолданылды: ғылыми әдебиеттерге шолу және контент-талдау, сауалнама әдісі. Ғылыми әдебиеттерге шолу арқылы дарынды оқушыларды оқытудағы инновациялық педагогикалық технологияларға қатысты зерттеулер талданып, ал сауалнама арқылы мұғалімдердің дарынды оқушыларды анықтау, дамыту және инновациялық технологияларды қолдану жөніндегі пікірлері талданды. Сауалнама нәтижесінде дарынды оқушыларды анықтауда (Сурет 1.) диагностикалық тест пен

практикалық тапсырмалар тиімді деп көрсеткен. Қабілетті дамытуда жеке оқу траекториясы (Сурет 2.), оқушы қабілетін анықтауда мұғалімнің үлесі басым болды (Сурет 3.). дарынды оқушылармен жұмыс барысында Peer Instruction және ChatGPT сияқты заманауи технологияларды қолдануға қажеттілігі жоғары екені (Сурет 4) анықталды.

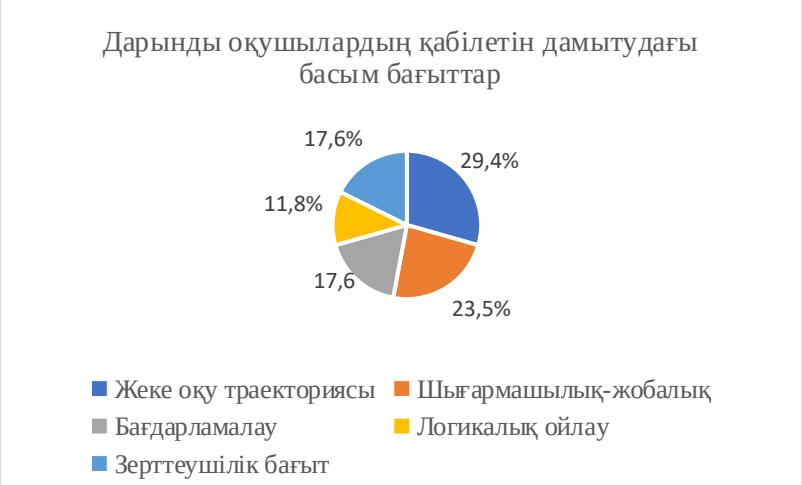
Кесте 1. Зерттеу барысында қолданылған сауалнама сұрақтары

№	Сауалнама	Зерттеу мақсаты
1	Сіздің пікіріңізше дарынды оқушыларды анықтауда қандай әдіс-тәсіл тиімді ?	Мұғалімдердің инновациялық технологияларды қабылдау дайындығын анықтау
2	Дарынды оқушылардың қабілетін дамытуда қай бағыт басым деп ойлайсыз?	Кәсіби және әдістемелік қажеттіліктерді анықтау
3	Оқушының қабілетін анықтауда кімнің үлесі басым деп санайсыз?	Дарындылықты диагностикалаудың тиімді тәсілдерін анықтау
4	Сабақта интерактивті оқыту әдістері (Case Study, PBL, Peer Instruction, Think–Pair–Share) қолданыла ма?	оқушылардың белсенді ойлауын, талдау, пікірталас, мәселе шешу және өз бетінше шешім қабылдау дағдыларын дамыту.



Респонденттер тиімді әдіс ретінде диагностикалық тест пен практикалық тапсырманы көрсеткен.

Сурет 1. Дарындылықты анықтау әдістері бойынша сауалнама нәтижесі



Сауалнама нәтижесі бойынша жеке оқу траекториясы басым бағыт деп танылған. Респонденттер шығармашылық-жобалық бағытты да маңызды деп көрсеткен.

Сурет 2. Дарындылықты дамытудағы басым бағыттарды анықтау бойынша сауалнама нәтижесі

Респонденттердің оқушы бойындағы дарындылық пен әлеуетті айқындауға әсер ететін негізгі субъектілер бойынша берген бағалау нәтижелері көрсетілген (3-сурет).

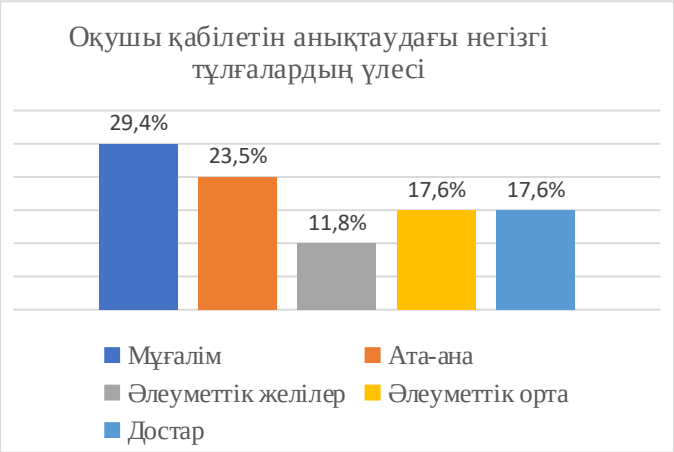
Мұғалімдер (29,4%): Оқушының академиялық және шығармашылық икемділігін білім беру үдерісінде алғаш болып таниды.

Ата-аналар (23,5%): Баланың жеке қасиеттері мен қызығушылықтарын алғашқы даму кезеңінде байқайтын екінші негізгі субъект ретінде анықталды.

Әлеуметтік орта (317,6%): Оқушының сыртқы ортамен өзара әрекеттесуі арқылы қабілеттерінің ашылуына ықпал етеді.

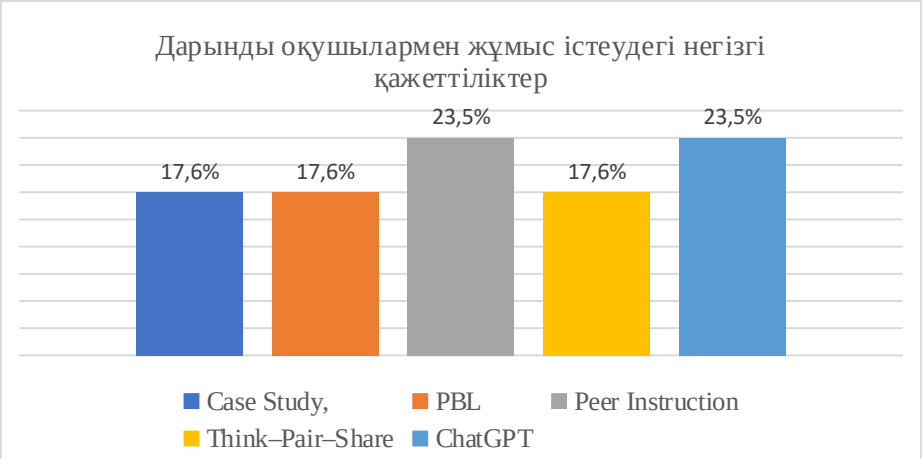
Достар (17,6%): Құрдастар арасындағы бәсекелестік пен ортақ қызығушылықтар оқушының бейімділігін көрсетуге мүмкіндік береді.

Әлеуметтік желілер (11,8%): Ақпараттық кеңістік оқушының қызығушылығын арттырғанымен, қабілетті айқындауда салыстырмалы түрде төмен ықпал етеді.



Сурет 3. Дарындылықты анықтауға әсер ететін субъектілер бойынша сауалнама нәтижесі

Дарынды оқушылармен жұмыс істеуде әдіс-тәсілдер мен құралдардың қолданылу жиілігі немесе басымдылығын анықтау нәтижесі 4-суретте көрсетілген.



Сурет 4. Педагогикалық әдістердің ролін анықтау бойынша сауалнама нәтижесі

Жоғары көрсеткішке ие педагогикалық құралдар: Peer Instruction (Құрдастардың өзара үйретуі) мен ChatGPT ең жоғары мәнге ие болып тұр. Бұл дарынды оқушылардың өзара пікірталас жүргізу арқылы білім алуы және жасанды интеллект құралдарын өз бетінше ізденуге белсенді қолданатынын көрсетеді.

Тұрақты қолданыстағы әдістер: Case Study, PBL (Жобаға бағытталған оқыту) және Think–Pair–Share әдістері біркелкі, орташа деңгейді көрсетіп тұр. Бұл интерактивті оқыту форматының дарынды оқушылардың сыни ойлауын дамытуда тұрақты ролі бар екенін білдіреді.

Қорытынды

Дарынды оқушылардың әлеуетін толық ашу үшін дәстүрлі диагностикалық әдістерді инновациялық педагогикалық технологиялармен ұштастыру қажет. Жеке оқу траекториясы, жобалық оқыту, бағдарламалау, белсенді оқыту әдістері және жасанды интеллект құралдары оқушылардың шығармашылық, сыни ойлау, зерттеушілік және

проблеманы шешу дағдыларын дамытуға мүмкіндік береді. Технологияны дайын шешім ретінде емес, оқушының дербестігі мен белсенділігін арттыратын педагогикалық құрал ретінде қолдануын көрсетті.

Әдебиеттер

1. Aydın, M., & Yurdugül, H. (2024). Developing a Curriculum Framework of Artificial Intelligence Teaching for Gifted Students. Kastamonu Education Journal, 32(1), 14–37. https://doi.org/10.24106/kefdergi.1426429
2. Avcu, Y. E., & Ayverdi, L. (2020). Examination of the computer programming self-efficacy's prediction towards the computational thinking skills of the gifted and talented students. International Journal of Educational Methodology, 6(2), 259–270. https://doi.org/10.12973/ijem.6.2.259
3. Kirmit, Ş., Dönmez, İ., & Çataltaş, H. E. (2018). Üstün yetenekli öğrencilerin bilgisayarca düşünme becerilerinin incelenmesi. Journal of STEAM Education, 1(2), 17–26.
4. Kırmit D., Dönmez O., & Çataltaş, B. (2018). Üstün yetenekli öğrencilerin bilgisayarca düşünme becerilerinin incelenmesi. Bilim, Teknoloji, Mühendislik, Matematik ve Sanat (J-STEAM) Eğitimi Dergisi, 1(2), 17–26. https://doi.org/10.17583/rise.16879
5. Sternberg, R. J. (2009). Assessment of gifted students for identification purposes: New techniques for a new millennium. *Learning and Individual Differences, 20*(4), 327–336. https://doi.org/10.1016/j.lindif.2009.08.003
6. Worrell F. C., Subotnik R. F., Olszewski-Kubilius P., & Dixson D. D. (2019). Gifted Students. Annual Review of Psychology, 70, 551–576. https://doi.org/10.1146/annurev-psych-010418-102846
7. Toplu, Ş., Ekinci, R., & Köklü, O. (2025). Designing and implementing a data science module for gifted middle school students. Journal of Pedagogical Research, 9(5), 324–346. https://doi.org/10.33902/JPR.202535630
8. Alqahtani, R., & Alqahtani, M. A. (2023). Heterogeneity across Australian ICT policies for education of gifted students. Heliyon, 9, e19103. https://doi.org/10.1016/j.heliyon.2023.e19103
9. Esparza, J., Shumow, L., & Schmidt, J. A. (2014). Growth mindset of gifted seventh grade students in science. NCSSSMST Journal, Spring 2014.
10. Мадьярова Г., Орынбаева А., Жалгасова А., Култан Я. и Сегизбаева Р. 2025. Подготовка преподавателей информации к работе с одаренными учащимися: современные методы и технологии. Вестник Абай КазНПУ. Серия физико-математических наук . 92, 4 (дек. 2025), 283–295. https://doi.org/10.51889/2959-5894.2025.92.4.027.

Өтініш

№	ФИО	Место работы авторов, должность;	Форма участия - очная или заочная;	Название доклада;	Предпочтительная форма доклада: пленарный, секционный, стендовый;	E-mail. Окончательное решение по форме доклада Оргкомитет оставляет за собой, рукописи не возвращаются.
1	Мадьярова Гульнар Атыхановна	Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Компьютерлік	(сырттай) Заочная	Дарынды оқушылармен жұмыс істеуде инновациялық педагогикалық технологиял	секционный	Madyarova-gulnar@mail.ru aizhan.sansyzbaevna@mail.ru

		ғылымдар кафедрасының асисстент профессоры		арды қолдану		
2	Орынбаева Айжан Сансызбаевна	Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Информатика педагогтерін даярлау докторанты				

ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ СҮТ БЕЗІ АУРУЛАРЫН ЕРТЕ АНЫҚТАУДЫҢ ИНТЕЛЛЕКТУАЛДЫ МОДЕЛЬДЕРІН ӘЗІРЛЕУ

Максутова Ш.У.
e-mail: shynar_04@mail.ru
Әл-Фараби атындағы Қазақ ұлттық университеті

АҢДАТПА

Мақалада жасанды интеллект пен үлкен деректер технологиялары негізінде сүт безі ауруларын ерте анықтаудың интеллектуалды моделін әзірлеу мәселесі қарастырылады. Маммографиялық кескіндерді алдын ала өңдеу, сегментациялау, белгілерді автоматты түрде шығару және жіктеу әдістері талданды. CNN және Transformer негізіндегі терең оқыту модельдерін қолдану мүмкіндіктері қарастырылды. Жарияланған жүйелі шолулар мен мета-талдаулардың нәтижелері салыстырылып, AI модельдерінің диагностикалық көрсеткіштері бағаланды. Ұсынылатын модель дәрігердің шешімін алмастырмай, диагностикалық шешім қабылдауды қолдауға бағытталған. Нәтижелер AI технологияларының сүт безі ауруларын ерте анықтауда перспективалы екенін көрсетеді.

Түйін сөздер: жасанды интеллект, машиналық оқыту, терең оқыту, үлкен деректер, сүт безі обыры, маммография, CNN, Transformer, Explainable AI, интеллектуалды жүйе.

Сүт безі обыры қазіргі таңда әлемдегі денсаулық сақтау жүйелері үшін маңызды медициналық мәселелердің бірі болып табылады. Дүниежүзілік денсаулық сақтау ұйымының 2026 жылғы мәліметі бойынша, 2024 жылы әлемде шамамен 2,4 млн әйелге сүт безі обыры диагнозы қойылып, 694 мың адам осы аурудан қайтыс болған. Сүт безі обыры 186 елдің 164-інде әйелдер арасында ең жиі кездесетін қатерлі ісік түрі болған.

Ауруды ерте анықтау пациенттің әрі қарайғы емдеу тактикасын таңдауда және аурудың ауыр салдарын азайтуда маңызды рөл атқарады. Дүниежүзілік денсаулық сақтау ұйымы (ДДҰ) ерте және уақтылы диагностиканы сүт безі обырының ауыртпалығын төмендетудің негізгі компоненттерінің бірі ретінде қарастырады.

Қазіргі кезде сүт безі ауруларын диагностикалауда маммография, ультрадыбыстық зерттеу және цифрлық томосинтез кеңінен пайдаланылады. Алайда медициналық кескіндер көлемінің ұлғаюы радиологтардың жұмыс жүктемесін арттырып, диагностикалық процесті автоматтандыру қажеттілігін күшейтіп отыр.

Осы мәселені шешудің перспективалы бағыттарының бірі — жасанды интеллектті пайдалану. Машиналық оқыту алгоритмдері үлкен көлемдегі медициналық деректерден заңдылықтарды анықтауға мүмкіндік береді. Ал терең оқыту, әсіресе конволюциялық нейрондық желілер (CNN), медициналық кескіндерден күрделі визуалды белгілерді автоматты түрде алуға қабілетті.

2022 жылғы жүйелі шолу мен мета-талдауға 32 зерттеу және 23 804 маммографиялық кескін енгізілген. Үш машиналық оқыту әдісін біріктіргенде жиынтық сезімталдық 91,4%, спецификалық 91,6%, AUC 0,945 болған. CNN модельдері үшін сәйкес көрсеткіштер 96,1%, 95,0% және 0,974 деңгейінде болған. Авторлар нәтижелер перспективалы болғанымен, зерттеулердің едәуір бөлігі ретроспективті болғандықтан, қатаң проспективті зерттеулер қажет екенін атап көрсеткен.

Сондықтан сүт безі ауруларын ерте анықтауға арналған AI жүйесін әзірлеу тек жоғары дәлдікті модель құрумен шектелмеуі тиіс. Деректер сапасы, модельдің түсіндірмелілігі, сыртқы валидация, жалған оң және жалған теріс нәтижелер, сондай-ақ медициналық жұмыс процесіне интеграциялау мәселелері қатар қарастырылуы қажет.

Зерттеудің мақсаты. Жасанды интеллект және үлкен деректер технологияларын пайдалана отырып, сүт безі ауруларын ерте анықтауға арналған интеллектуалды диагностикалық модельдің құрылымын әзірлеу және оның негізгі компоненттерінің тиімділігін ғылыми әдебиеттер негізінде бағалау.

Зерттеу міндеттері. Зерттеу мақсатына жету үшін келесі міндеттер қойылды:
1. сүт безі ауруларын диагностикалауда қолданылатын медициналық деректерді зерттеу;
2. маммографиялық кескіндерді алдын ала өңдеу әдістерін анықтау;
3. медициналық кескіндерден белгілерді автоматты түрде шығару әдістерін талдау;
4. CNN және Transformer архитектураларын қолдану мүмкіндіктерін қарастыру;
5. сүт безі түзілістерін жіктейтін интеллектуалды модельдің архитектурасын ұсыну;
6. модельдің тиімділігін бағалау критерийлерін анықтау;
7. Explainable AI әдістерін интеллектуалды жүйеге енгізу;
8. AI моделін клиникалық жұмыс процесіне енгізудің негізгі талаптарын анықтау.

Зерттеу материалдары мен әдістері. Зерттеуде сүт безі ауруларын диагностикалауда жасанды интеллектті қолдануға арналған ғылыми жарияланымдарға жүйелі талдау жасалды. Негізгі назар маммографиялық кескіндерді өңдейтін машиналық оқыту және терең оқыту модельдеріне аударылды.

Медициналық деректер. Ұсынылатын жүйенің негізгі деректер көздері:
- цифрлық маммография;
- цифрлық сүт безі томосинтезі;
- ультрадыбыстық кескіндер;
- пациенттің жасы;
- сүт безінің тығыздығы;
- бұрынғы зерттеулер;
- радиологиялық қорытындылар;
- биопсия және гистологиялық нәтижелер.

Мұндай деректерді біріктіру мультимодальды интеллектуалды жүйе құруға мүмкіндік береді.

Деректерді алдын ала өңдеу. Медициналық кескіндерді модельге енгізбес бұрын бірнеше өңдеу кезеңі орындалады:

1. Деректерді тазалау. Қайталанатын, сапасы төмен және техникалық ақауы бар кескіндер анықталады.

2. Нормализация. Кескіндердің интенсивтілік диапазоны стандартталады.

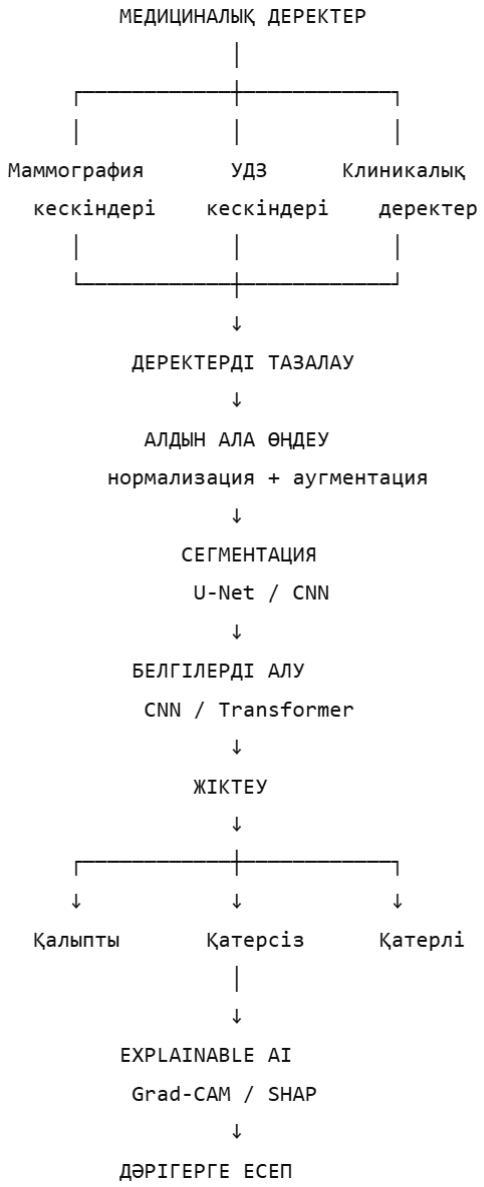
3. Өлшемді стандарттау. Әртүрлі құрылғылардан алынған кескіндер бірдей форматқа келтіріледі.

4. Аугментация. Кескінді айналдыру, масштабтау, контрастты өзгерту сияқты әдістер пайдаланылады.

5. Сегментация. Сүт безі және күмәнді зақымдану аймақтары бөлініп алынады.

Ұсынылатын интеллектуалды модель

Ұсынылатын интеллектуалды жүйе бес негізгі кезеңнен тұрады (1-сурет):
1. деректерді қабылдау және тазалау;
2. алдын ала өңдеу;
3. зақымдану аймағын сегментациялау;
4. терең нейрондық желі арқылы белгілерді алу және жіктеу;
5. нәтижені түсіндіру және дәрігерге ұсыну.



1-сурет. Ұсынылатын интеллектуалды жүйенің архитектурасы

Модельдің ерекшелігі — диагностикалық нәтижені тек «қатерлі» немесе «қатерсіз» деп көрсету емес, шешімге әсер еткен кескін аймағын да көрсету.

Модельді бағалау критерийлері

Медициналық AI жүйесін бағалау кезінде бірнеше көрсеткіш қатар қолданылады.

Сезімталдық:

Sensitivity=\frac{TP}{TP+FN}

Сезімталдық ауруы бар пациенттердің қанша бөлігін модель дұрыс анықтағанын көрсетеді.

Спецификалық:

Specificity=\frac{TN}{TN+FP}

Спецификалық сау жағдайлардың қанша бөлігінің дұрыс танылғанын сипаттайды.

Precision:

Precision=\frac{TP}{TP+FP}

F1-score:

F1=2\frac{Precision\timesRecall}{Precision+Recall}

ROC-AUC: ROC-AUC моделі қатерлі және қатерсіз жағдайларды қаншалықты жақсы ажырататынын көрсетеді.

Медициналық скрининг жағдайында сезімталдықтың жоғары болуы ерекше маңызды, өйткені жалған теріс нәтиже ауруды уақытында анықтамауға алып келуі мүмкін.

Зерттеу нәтижелері

Машиналық оқыту модельдерінің диагностикалық көрсеткіштері (1-кесте)

Wang және әріптестері жүргізген жүйелі шолу мен мета-талдауда 32 зерттеу және 23 804 маммографиялық кескін талданған. Үш негізгі машиналық оқыту әдісі — CNN, ANN және SVM — салыстырылған.

1-кесте. Маммографиядағы машиналық оқыту әдістерінің диагностикалық көрсеткіштері

Модель	Сезімталдық	Спецификалық	AUC
Барлық ML әдістері	91,4%	91,6%	0,945
CNN	96,1%	95,0%	0,974
ANN	83,7%	89,4%	0,881
SVM	88,9%	84,3%	0,913

Дереккөз: Wang et al., жүйелі шолу және мета-талдау.

Модельдердің сезімталдығы мен спецификалығы

Маммография бойынша жарияланған мета-талдауда көрсетілген сезімталдық және спецификалық көрсеткіштер (2-сурет).



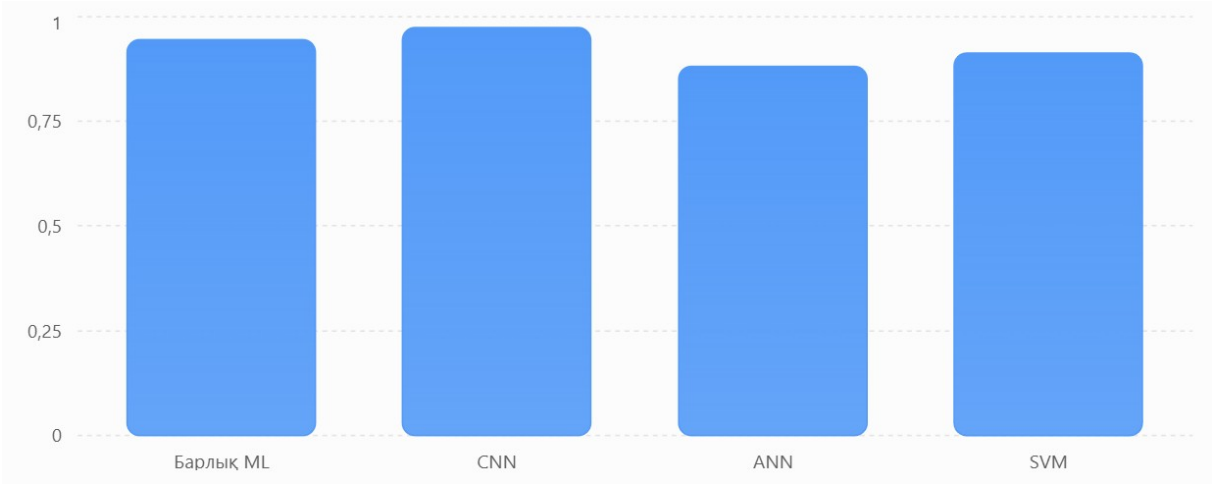
2-сурет. Модельдердің сезімталдығы мен спецификалығын салыстыру

Графиктен CNN моделінің қарастырылған үш әдістің ішінде жоғары сезімталдық және спецификалық көрсеткіштерге ие болғаны көрінеді. CNN үшін сезімталдық 96,1%, ал спецификалық 95,0% болған.

ROC-AUC көрсеткіштерін салыстыру

AUC көрсеткіші диагностикалық модельдің екі классты ажырату қабілетін бағалауда маңызды.

Әртүрлі машиналық оқыту әдістерінің жарияланған жиынтық AUC көрсеткіштері (3-сурет)



3-сурет – Машиналық оқыту модельдерінің ROC-AUC көрсеткіштері

CNN үшін AUC 0,974 деңгейінде болған. Бұл көрсеткіш зерттелген ретроспективті деректер жағдайында модельдің қатерлі және қатерсіз жағдайларды ажырату қабілетінің жоғары болғанын көрсетеді. Дегенмен бұл нәтижені нақты клиникалық практикаға тікелей көшіруге болмайды, себебі мета-талдауға енгізілген зерттеулердің басым бөлігі ретроспективті болған.

AI триажының радиолог жұмыс жүктемесіне әсері

AI технологиясының тағы бір маңызды қолданылу саласы — маммографиялық зерттеулерді триаждау.

Xavier және әріптестерінің 2024 жылғы мета-талдауына 156 852 маммографиялық зерттеу енгізілген. Талдау нәтижесінде белгілі бір AI алгоритмінің параметрлерінде радиологтардың жұмыс жүктемесі теориялық түрде 68,3%-ға азайған, ал сезімталдық 93,1%, спецификалық 68,7% болған (2-кесте).

2-кесте. AI триажының жарияланған нәтижелері

Көрсеткіш	Нәтиже
Талдауға енгізілген зерттеу саны	3
Маммографиялық тексерулер	156 852
Теориялық жұмыс жүктемесінің төмендеуі	68,3%
Сезімталдық	93,1%
Спецификалық	68,7%

Дереккөз: Xavier et al., 2024.

AI және радиологиялық диагностиканы салыстыру

Standalone AI жүйелерінің диагностикалық өнімділігін зерттеген 2023 жылғы жүйелі шолу мен мета-талдауда цифрлық маммография бойынша reader-study зерттеулерінде AI үшін жинақталған AUC 0,87, ал радиологтар үшін 0,81 болған. Historical cohort studies жағдайында AI AUC 0,89, радиологтар AUC 0,96 көрсеткен (3-кесте).

3-кесте. AI және радиологтардың AUC көрсеткіштері

Зерттеу түрі	AI AUC	Радиологтар AUC
Reader studies	0,87	0,81
Historical cohort studies	0,89	0,96

Бұл нәтижелер AI моделінің тиімділігі зерттеу дизайнына және қолдану жағдайына тәуелді екенін көрсетеді. Сондықтан AI жүйесін клиникалық практикаға енгізу кезінде тек бір зерттеудің AUC көрсеткішіне сүйену жеткіліксіз.

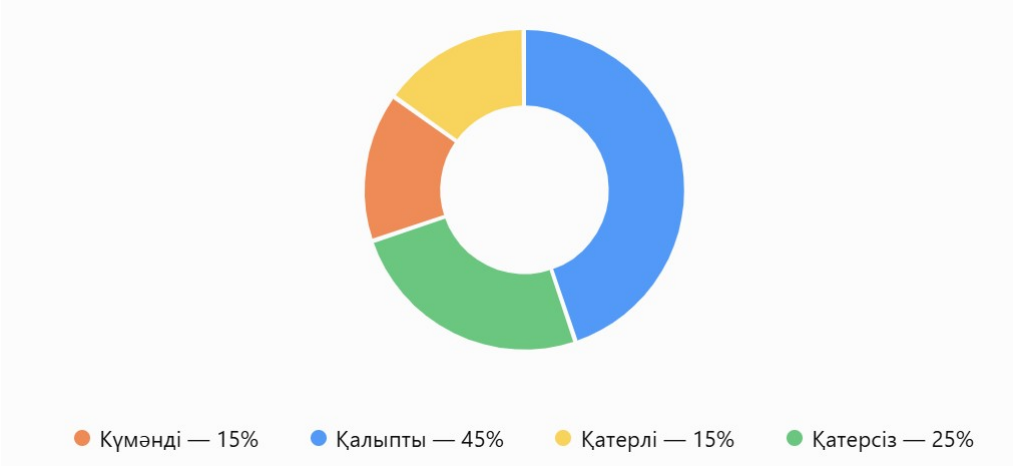
Интеллектуалды жүйенің нәтижелерін визуализациялау

Интеллектуалды жүйенің диагностикалық нәтижелерді визуализациялау үлгісі

Ұсынылатын жүйеде диагностикалық нәтиже төрт деңгейде көрсетілуі мүмкін:

- қалыпты;
- қатерсіз;
- күмәнді;
- қатерлі.

Төмендегі диаграмма жүйенің интерфейсін визуализациялау үшін ғана берілген концептуалдық мысал. Пайыздар нақты пациенттер жиынтығынан алынған эксперименттік деректер емес (4-сурет).



4-сурет. Интеллектуалды диагностикалық жүйенің нәтижелерді төрт санатқа бөлу үлгісі

Мұндай визуализация дәрігерге үлкен көлемдегі зерттеу нәтижелерін жылдам түсінуге мүмкіндік береді. Нақты жүйеде бұл пайыздар модельдің нақты тестілік немесе клиникалық деректер жиынтығынан есептелуі тиіс.

Түсіндірілетін жасанды интеллект

Медициналық диагностикада AI моделінің нәтижесін түсіндіру маңызды. Терең нейрондық желілер көбіне «қара жәшік» ретінде қабылданады. Сондықтан дәрігер модельдің қандай белгіге сүйеніп шешім қабылдағанын көре алуы қажет.

Ұсынылатын жүйеде Grad-CAM және SHAP әдістерін қолдану қарастырылады.

Grad-CAM көмегімен модель назар аударған кескін аймағын жылу картасы түрінде көрсетуге болады. SHAP әдісі енгізілген белгілердің соңғы шешімге қосқан үлесін бағалауға мүмкіндік береді.

Мысалы, жүйе күмәнді аймақты анықтаған жағдайда дәрігерге:

Қатерлілік ықтималдығы → модель назар аударған аймақ → сенімділік деңгейі → қосымша тексеру қажеттілігі түрінде ақпарат беріледі.

Бұл тәсіл AI жүйесін дәрігердің шешімін алмастыратын құрал ретінде емес, шешім қабылдауды қолдайтын интеллектуалды жүйе ретінде қолдануға мүмкіндік береді.

Үлкен деректер және деректер ғылымының рөлі

Сүт безі ауруларын ерте анықтауға арналған AI жүйесінің тиімділігі тек нейрондық желінің архитектурасына емес, қолданылатын деректердің сапасына да тәуелді.

Үлкен медициналық деректер келесі сипаттармен ерекшеленеді:

Үлкен деректер сипаттамасы	Сүт безі диагностикасындағы көрінісі
Volume	Мыңдаған және миллиондаған кескін
Variety	Маммография, УДЗ, DBT, клиникалық ақпарат
Velocity	Жаңа медициналық зерттеулердің үздіксіз түсуі
Veracity	Белгілеулердің және медициналық ақпараттың сенімділігі
Value	Диагностика үшін пайдалы заңдылықтарды анықтау

Деректер ғылымы деректерді жинау, тазалау, нормализациялау, статистикалық талдау, визуализация және машиналық оқыту модельдерін әзірлеу кезеңдерінің барлығында қолданылады.

Сонымен бірге пациенттердің жеке медициналық ақпаратын қорғау қажет. Деректерді анонимдеу, қолжетімділікті шектеу және қауіпсіз сақтау медициналық AI жүйесінің міндетті компоненттері болуы тиіс.

Модельді клиникалық валидациялау

AI моделін нақты клиникалық практикаға енгізбес бұрын бірнеше деңгейлі валидация қажет.

Бірінші деңгей — ішкі валидация. Модель оқытуға пайдаланылған деректерге ұқсас тестілік деректерде бағаланады.

Екінші деңгей — сыртқы валидация. Модель басқа медициналық ұйымнан алынған деректерде тексеріледі.

Үшінші деңгей — көпорталықты валидация. Әртүрлі аппараттар мен пациенттер популяциясында модельдің тұрақтылығы бағаланады.

Төртінші деңгей — проспективті зерттеу. AI жүйесінің нақты клиникалық жұмыс процесіндегі нәтижесі бағаланады.

Бұл талаптар маңызды, себебі жарияланған зерттеулер арасында ретроспективті жұмыстардың үлесі жоғары.

2026 жылғы жүйелі шолуда сүт безі скринингіндегі AI бойынша 31 зерттеу және екі миллионнан астам скринингтік тексеру қамтылған.

Талқылау. Зерттеу нәтижелері жасанды интеллекттің сүт безі ауруларын ерте анықтауда маңызды технологиялық әлеуетке ие екенін көрсетеді.

Ең жоғары көрсеткіштердің бірі CNN негізіндегі модельдерде байқалады: жарияланған мета-талдауда CNN үшін сезімталдық 96,1%, спецификалық 95,0%, AUC 0,974 болған.

Сонымен қатар AI-ды радиолог жұмысын толық алмастыратын технология ретінде қарастыруға болмайды. Әртүрлі зерттеулердегі нәтижелер модельдің нақты қолдану ортасына тәуелді екенін көрсетеді. Standalone AI бойынша зерттеулерде AI өнімділігі reader study және historical cohort жағдайларында әртүрлі болған.

AI триажының тағы бір перспективалы бағыты — радиологтардың жұмыс жүктемесін азайту. 156 852 зерттеуді қамтыған мета-талдауда белгілі бір алгоритм жағдайында теориялық жұмыс жүктемесінің 68,3%-ға төмендеуі және сезімталдықтың 93,1%-ға жетуі көрсетілген.

Осыған байланысты ең орынды архитектуралардың бірі — адам + AI моделі. Бұл жағдайда AI күмәнді кескіндерді анықтап, басымдық береді, ал соңғы клиникалық шешімді дәрігер қабылдайды.

Ұсынылатын модельдің келесі даму кезеңдерінде мультимодальды деректерді пайдалану маңызды. Яғни маммография кескінін пациенттің клиникалық ақпараттарымен және басқа бейнелеу әдістерінің нәтижелерімен біріктіру модельдің диагностикалық мүмкіндігін кеңейтуі мүмкін.

Зерттеудің ғылыми және практикалық маңызы

Зерттеудің ғылыми маңызы — медициналық кескіндерді талдау үшін жасанды интеллект, терең оқыту, үлкен деректер және түсіндірілетін AI технологияларын бір интеллектуалды архитектураға біріктіру мүмкіндігін негіздеуде.

Практикалық маңызы — ұсынылатын жүйе:

- маммографиялық кескіндерді автоматты талдауға;
- күмәнді аймақтарды анықтауға;
- зерттеулерді қауіп деңгейі бойынша сұрыптауға;
- радиологтың жұмыс процесін қолдауға;
- диагностикалық ақпаратты визуалды түрде ұсынуға;
- модель шешімін түсіндіруге мүмкіндік береді.

Бұл ретте жүйе медициналық маманның кәсіби шешімін алмастырмайды.

Қорытынды. Зерттеу барысында жасанды интеллект, машиналық оқыту, терең оқыту және үлкен деректер технологиялары негізінде сүт безі ауруларын ерте анықтауға арналған интеллектуалды модельдің архитектурасы ұсынылды.

Ұсынылған модель медициналық деректерді қабылдау, тазалау, алдын ала өңдеу, сегментациялау, CNN/Transformer арқылы белгілерді шығару, жіктеу және Explainable AI арқылы нәтижені түсіндіру кезеңдерінен тұрады.

Ғылыми әдебиеттерді талдау CNN негізіндегі модельдердің маммографиялық кескіндерді талдауда жоғары көрсеткіштерге қол жеткізе алатынын көрсетті. Жарияланған мета-талдауда CNN үшін сезімталдық 96,1%, спецификалық 95,0% және AUC 0,974 болған.

AI технологиясын триаж үшін пайдалану да перспективалы бағыт болып табылады. 156 852 маммографиялық тексеруді қамтыған мета-талдауда белгілі бір алгоритм параметрлері кезінде теориялық радиолог жұмыс жүктемесінің 68,3%-ға төмендеуі және 93,1% сезімталдық көрсетілген.

Сонымен бірге AI моделінің жоғары техникалық көрсеткіші оның автоматты түрде клиникалық қолдануға дайын екенін білдірмейді. Сыртқы және проспективті валидация, деректердің әртүрлілігі, модельдің түсіндірмелілігі және медициналық жұмыс процесіне қауіпсіз интеграциялау міндетті түрде қарастырылуы тиіс.

Осылайша, жасанды интеллектке негізделген интеллектуалды жүйелер сүт безі ауруларын ерте анықтау, диагностикалық процесті қолдау және медициналық ресурстарды тиімді пайдалану үшін перспективалы бағыт болып табылады. Ұсынылған модельді одан әрі дамытуда мультимодальды деректер, Transformer архитектуралары, Explainable AI және көпорталықты клиникалық валидация негізгі зерттеу бағыттары ретінде қарастырылады.

ӘДЕБИЕТТЕР

1. World Health Organization. Breast cancer. WHO Fact Sheet, 3 July 2026.
2. World Health Organization. Cancer. WHO Fact Sheet, 2026.
3. International Agency for Research on Cancer. Breast cancer. Global Cancer Observatory. IARC/WHO, 2026.
4. Wang M., et al. Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis. *Breast Cancer Research and Treatment.* 2022.
5. Yoon J.H., Strand F., Baltzer P.A.T., et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. *Radiology.* 2023.
6. Xavier D., Miyawaki I., Jorge C.A.C., et al. Artificial intelligence for triaging of breast cancer screening mammograms and workload reduction: A meta-analysis of a deep learning software. *Journal of Medical Screening.* 2024;31(3):157–165.
7. Performance of artificial intelligence in breast cancer screening programmes: a systematic review. 2026.

8. Raya-Povedano J.L., et al. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. *Radiology.* 2021;300(1):57–65.
9. World Health Organization. Global Breast Cancer Initiative. WHO.
10. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. WHO.

УДК 004.9:159.9

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ЦВЕТОВОГО ПРИОРИТИЗАЦИИ И ПЛАНИРОВАНИЯ ВРЕМЕНИ ДЛЯ БОРЬБЫ С ЦИФРОВОЙ АПАТИЕЙ

Г.З. Зиятбекова[1,2], Б.А. Абдуали[1], С. Жаныбекулы[2], А.В. Дудин[2], Т.А. Андронова[2]

[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
[2]Алматинский технологический университет, Алматы, Казахстан

Аннотация. *В данной статье рассматривается проблема рационального управления личным временем и бизнес-процессами в условиях цифровой гиперстимуляции. Вводится феномен «цифровой апатии» (также известной как бытовая деструктивная прокрастинация) и исследуется метод динамического планирования, основанный на жесткой цветовой приоритезации задач и интеграции алгоритмов искусственного интеллекта. Описывается концепция приложения-сервиса, автоматизирующего перестройку повседневного графика при изменениях внешних обстоятельств и оптимизирующего совместный досуг. Результаты работы показывают снижение когнитивной нагрузки и рост продуктивного использования ресурсов пользователя.*

Ключевые слова: *цифровая апатия, интеллектуальная система, оптимизация, пользовательский опыт, повседневная польза.*

Введение. Современная парадигма развития информационного общества характеризуется экспоненциальным ростом объемов поступательной информации и сужением временных рамок, отведенных человеку на адаптацию к меняющимся условиям среды. Процессы глобальной цифровизации сформировали феномен постоянной включенности (культура «always-on»). Однако доступность специализированных приложений для тайм-менеджмента и цифровых календарей не привела к росту удовлетворенности качеством жизни. Напротив, в практике все чаще фиксируется рост когнитивной перегрузки, выгорания и специфических форм деструктивного поведения.

Ключевым фактором, дестабилизирующим индивидуальную продуктивность, выступает *синдром усталости от принятия решений* (*decision fatigue*). В течение дня человек совершает тысячи микровыборов. К моменту завершения обязательной трудовой или учебной программы волевой ресурс нервной системы оказывается истощенным. В таком состоянии необходимость самостоятельно планировать досуг, выбирая между чтением, спортом, обучением или бытовыми делами, воспринимается когнитивной системой как дополнительная нежелательная нагрузка [1].

Основная часть. Именно в этой точке возникает патологический паттерн поведения, определяемый в рамках настоящего исследования как *«цифровая апатия»* (или бытовая деструктивная прокрастинация) (Таблица 1).

Таблица 1. Механизм возникновения цифровой апатии при истощении когнитивного ресурса

Внешний стресс / Усталость после работы
↓
Истощение волевого ресурса (Decision Fatigue / Усталость)
↓
Необходимость вручную составлять план на вечер или выбирать хобби
↓

Отказ от сложности -> Переход в компульсивный медиа скроллинг (Получение "дешевого" дофамина с 0% отдачей)

Феномен цифровой апатии существенно отличается от классического физиологического отдыха. Если полноценный отдых направлен на восстановление ресурса центральной нервной системы, то цифровая апатия представляет собой компульсивный ввод однотипных высокочастотных стимулов посредством бесконечного прокручивания лент коротких видеороликов (формата TikTok, Shorts, Reels). С точки зрения нейробиологии, данный процесс генерирует быстрый «дешевый» дофамин с нулевым полезным эффектом для развития личности, оставляя человека в состоянии эмоционального истощения и чувства вины за впустую потраченные часы [3].



В пр тает быть
абстрактн ременной
слот имее и вектор
целевого назначения. Фундаментальным отличающим фактором системы является детальная проработка четырех категорий, исключающая двойственность трактовки и обеспечивающая однозначную реакцию алгоритмов приложения при изменениях внешних условий [5].

Черная категория (Базовый восстановительный приоритет) представляет собой фундаментальный базис жизнедеятельности человека, предназначенный для физиологического восстановления, сохранения психоэмоционального здоровья и обеспечения необходимого уровня жизнеобеспечения организма. В Черный сектор закладываются физиологически обусловленные процессы: полноценный ночной и дневной сон (например, строго зафиксированный интервал с 00:00 до 07:00), обязательные утренние и вечерние гигиенические ритуалы, спокойные приемы пищи и периоды абсолютного личного покоя. Под последним понимается время, когда человек намеренно абстрагируется от социальных контактов, цифровых уведомлений и интеллектуальных нагрузок. Черная зона обладает абсолютным 100%-ным приоритетом и высшим уровнем жесткости. Данный интервал принципиально не может быть сокращен, сдвинут или перекрыт рабочими, академическими или бытовыми задачами. В традиционных системах планирования время сна и отдыха часто оставляют пустым фоном, что приводит к его систематическому вытеснению другими задачами. В нашей концепции Черная зона фиксируется в виде жестких графических блоков, за сдвиг которых система выдает предупреждения о рисках когнитивного истощения. Сохранность этого блока обеспечивает базовый ресурс центральной нервной системы, без которого повышение продуктивности в остальных сферах неизбежно приводит к выгоранию.

Красная категория (Жесткий внешний приоритет) консолидирует виды деятельности, выполнение которых регламентируется внешними социальными, профессиональными, академическими или институциональными правилами. В Красный сектор входят процессы с четкими временными рамками и внешними дедлайнами: рабочий график, производственные смены, очные и дистанционные академические занятия (например, пара в университете с 08:30 до 10:25), плановые записи к медицинским специалистам, бизнес-переговоры и официальные мероприятия. Красная

зона имеет высшую степень исполнительской обязательности, так как за ее нарушение следуют прямые внешние санкции в виде увольнения, академической задолженности или финансовых потерь. Однако, несмотря на внешнюю жесткость, она подвержена влиянию случайных факторов: преподаватель может отменить занятие, рабочий звонок завершиться раньше запланированного времени, а встреча быть перенесена. Поэтому Красная зона является главным генератором незапланированных временных окон. Когда событие отменяется, у человека внезапно высвобождается временной ресурс, который без автоматизированной системы планирования почти гарантированно поглощается цифровой апатией в ожидании следующего дела.

Желтая категория (Динамический переменчивый приоритет) охватывает важные сферы бытового комфорта, личного роста и поддержания необходимого порядка, которые не имеют строгого внешнего контроля. В Желтый сектор включаются плановые бытовые и хозяйственные задачи: уборка, закупка продуктов, приготовление пищи, административные вопросы и другие дела, которые необходимо выполнить, но время их выполнения может быть гибко изменено. Желтая зона выступает самым пластичным и адаптивным элементом системы. В классических календарях именно задачи из этой категории массово откладываются: когда человек устает после учебы или работы, у него не остается волевого ресурса заниматься бытовыми делами. Возникает пропуск задачи, за которым следует чувство вины, стресс и уход в цифровую апатию. В нашей системе Желтая зона оснащена механизмом мягкой адаптации. Если пользователь фиксирует снижение когнитивного ресурса или у него меняется график, алгоритмы приложения не помечают Желтую задачу как безусловно просроченную. Система автоматически переносит ее на наиболее благоприятный будущий слот, сохраняя общую структуру недели и освобождая пользователя от необходимости самостоятельно принимать это решение.

Зеленая категория (Свободный продуктивный приоритет) зарезервирована под свободный, ничем не занятый, но конструктивный и осмысленный досуг, выступающий главным инструментальным антагонистом цифровой апатии и компульсивного медиапотребления. В Зеленый сектор входит широкий спектр активностей: творческие хобби, музыка, моделирование, чтение художественной литературы, прогулки, настольные или компьютерные игры, социализация и живое общение с друзьями, семейный досуг, а также подвижные игры. Важно четко разграничивать Черную и Зеленую зоны. Если Черная зона — это пассивный покой и сон, необходимые для физиологической перезагрузки, то Зеленая зона представляет собой активную смену деятельности с нулевым когнитивным давлением. Зеленый сектор не требует принятия сложных решений и не несет оценочной нагрузки, но при этом задействует творческие и эмоциональные центры мозга. Главная задача Зеленой категории — заранее сформировать для пользователя привлекательный и готовый сценарий времяпрепровождения на момент, когда он завершает учебу или работу либо сталкивается с отменой планов. Без зафиксированной Зеленой зоны истощенный мозг выбирает путь наименьшего сопротивления в виде получения быстрого дофамина через бесконечный скроллинг коротких видеороликов. Зеленая зона создает структурированный барьер, превращая свободное время в полноценный и глубокий отдых.

Нейросетевая адаптация временных рамок является дополнительным уровнем автоматизации, объединяющим все четыре категории в единую систему. По задумке, алгоритмы программы не воспринимают заданные пользователем временные рамки как абсолютно точные значения, а способны анализировать реальное выполнение задач и выявлять возникающие погрешности. Система учитывает такие факторы, как время на дорогу до дома или учебного учреждения, задержки транспорта, непредвиденные обстоятельства, фактическую продолжительность выполнения предыдущей задачи и другие отклонения от первоначального плана. На основании накопленных данных

нейросетевые алгоритмы автоматически корректируют продолжительность и расположение последующих временных блоков, формируя более реалистичный график. Таким образом, пользователь задает общую структуру недели, а программа самостоятельно адаптирует ее к реальным условиям, постепенно снижая необходимость постоянного ручного редактирования расписания.

Практическое применение недельных шаблонов и режим принятия решений. Практическая реализация методологии четырехуровневой приоритезации строится на проектировании базовых недельных шаблонов. Вместо ежедневного формирования списка дел «с нуля», человек единоразово настраивает типовой шаблон своей недели, отражающий устойчивые жизненные ритмы. В шаблоне жестко зафиксированы Черная зона сна и отдыха, а также Красная зона работы или учебы. Желтые и Зеленые интервалы распределяются по остаточному принципу, формируя гибкий каркас дня [2].

Основная ценность использования базового шаблона заключается в кардинальном снижении когнитивной нагрузки на пользователя в моменты снижения волевого ресурса. Когда человек находится в состоянии апатии, усталости или высокой эмоциональной истощенности, ему категорически противопоказано принимать оперативные решения по выбору текущей деятельности. В этот момент приложение переходит в режим жесткого ведения по шаблону, где пользователь просто выполняет предписанный цветовой блок, либо позволяет алгоритмам автоматически сменить Желтый сектор на Зеленый без мучительного выбора конкретной задачи (Рисунок 2).

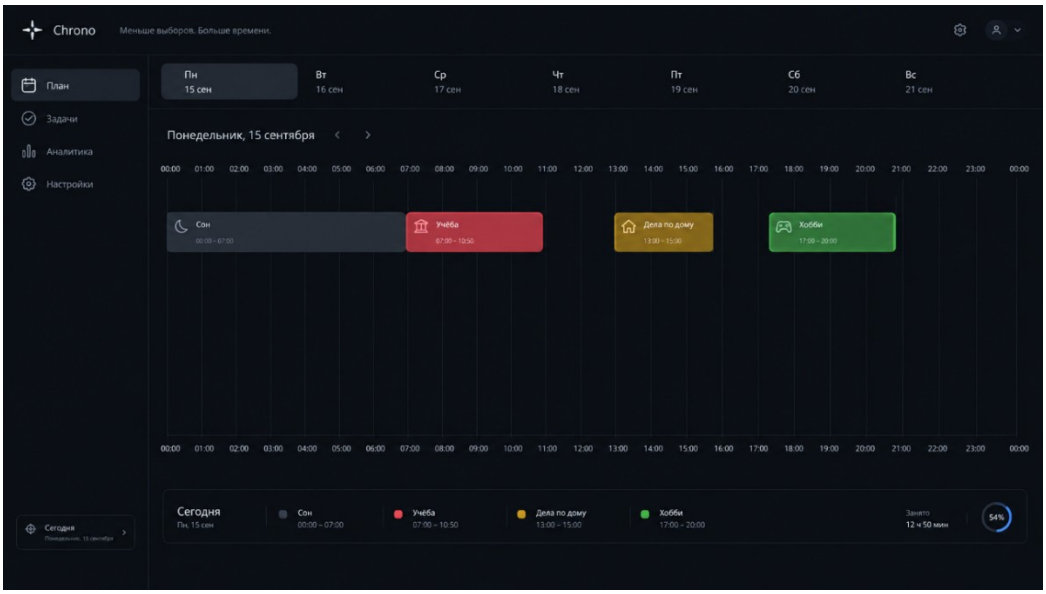


Рисунок 2. Цветовое кодирование задач в приложении

Интеграция искусственного интеллекта и алгоритмы динамической перестройки графика. Ключевым технологическим ядром проектируемого приложения-сервиса является встроенный модульный искусственный интеллект. Его главная задача — взять на себя рутинное оперативное перепланирование расписания при изменениях внешних обстоятельств, освобождая человека от постоянного принятия решений. При сбое в привычном распорядке дня (например, отмене лекции из Красной зоны или досрочном завершении рабочей встречи) ИИ-ассистент запускает алгоритм реструктуризации суточного графика. Он оценивает высвободившийся временной ресурс, текущее время суток и предполагаемое состояние пользователя. Если система предполагает высокий уровень усталости, освободившееся время не заполняется сложными задачами, а автоматически трансформируется в Зеленую зону. Ассистент подбирает готовый сценарий качественного отдыха, снижая необходимость самостоятельно принимать решение о занятии и риск ухода в компульсивный медиаскроллинг. Если же пользователь

находится в продуктивном состоянии, ИИ может перенести на освободившийся интервал задачи из Желтой зоны, освобождая последующий период для полноценного продуктивного досуга.

Модуль парной синхронизации для планирования совместного досуга. Отдельной важной функцией интеллектуального сервиса является модуль автоматической синхронизации графиков для пар и семейных союзов. На практике совместный отдых партнеров часто оказывается некачественным из-за того, что после тяжелого рабочего дня оба человека испытывают усталость от принятия решений и не могут согласовать формат вечера. В результате совместное время сводится к пассивному параллельному просмотру коротких видео или лент социальных сетей [4].

Интеллектуальный модуль решает эту проблему путем автоматического векторного наложения недельных шаблонов обоих партнеров. Алгоритм выявляет пересечения Зеленых и Желтых зон, сопоставляет географическое положение пользователей, анализирует их индивидуальные интересы и текущий уровень усталости. На основе этих данных ИИ самостоятельно формирует 2–3 конкретных и готовых варианта совместного досуга (например, прогулка по конкретному маршруту, посещение заведения или настольная игра). Это полностью устраняет стадию долгих споров и согласований, сохраняя психоэмоциональные ресурсы партнеров и повышая качество их совместной жизни.

Заключение. В данной работе рассмотрена проблема «цифровой апатии» и компульсивного медиаскроллинга, возникающих из-за усталости от принятия решений. Для её решения представлена система планирования на основе четырех цветовых категорий (Черная — восстановление, Красная — обязательства, Желтая — гибкий быт, Зеленая — продуктивный досуг), позволяющая однозначно распределить когнитивную нагрузку и защитить личное время. Внедрение этой цветовой структуры в сочетании с алгоритмами ИИ позволило автоматизировать перестройку графика, снизить волевое напряжение при сбоях в расписании и сократить время нецелевого медиапотребления. Работа необходима как практическая база для создания интеллектуальных сервисов, помогающих преодолеть выгорание и повысить качество личного и совместного отдыха.

Литература

1. Авдеев В.А., Соколова Е.М. Синдром эмоционального выгорания и методы когнитивно-поведенческой коррекции в условиях цифровизации труда // Вестник практической психологии образования, 2021. - Т. 18. - № 3. - С. 44–56.

2. Варламов И.П., Григорьев С.А. Алгоритмы динамической адаптации пользовательских графиков в интеллектуальных агентах персонального планирования // Труды Институт системного программирования РАН, 2023. - Т. 35. - Вып. 2. - С. 101–116.

3. Иванов И.И., Петров П.П. Психофизиология цифровой утомляемости и феномен клипового мышления в информационном обществе // Журнал высшей нервной деятельности, 2022. - Т. 45. - № 2. - С. 112–125.

4. Ковалева Н.В. Влияние социативно-медийных сервисов на когнитивную сферу молодежи: феномен бытовой прокрастинации // Вопросы психологии, 2020. - № 5. - С. 89–98.

5. Сидоров К.В. Автоматизация персонального тайм-менеджмента на основе адаптивных алгоритмов машинного обучения // Информационные системы и технологии, 2023. - № 1. - С. 45–58.

Ghost-CA-YOLO архитектурасы негізінде жол зақымдануларын нақты уақыт режимінде анықтау және бағалау

Самбетбаева А.К., Қажыбек А.М., Байкувеков М.Б.

Машиналық оқыту және терең оқыту әдістері жолдың зақымдануын анықтау және жөндеу жұмыстарында өте белсенді қолданылады. Машиналық оқыту жолдың зақымдануын анықтау үшін әртүрлі зерттеулердің деректерін жасауда қолданыла бастады. Машиналық оқыту қолданбаларының дамуы жолдың зақымдануын ерте анықтауға болатынын көрсетеді.

Кескінді өңдеу әдістерін қолдана отырып, жолдың зақымдануын анықтау бойынша зерттеулер белсенді түрде жүргізілуде, нәтижесінде анықтау дәлдігі айтарлықтай жоғары болады.

Жол төсеміндегі жарықтар мен шұңқырлардың қалыптасу механизмін зерттеу жүргізілді. Зерттеу барысында binding–pointing–crossword негіздемесі қолданылып, жарықтардың бағытталуы, байланысуы және қиылысу аймақтары талданды. Dataset жинау стратегиясы әзірленіп, авторлық деректер мен ашық dataset-терді біріктіру жоспарланды. Келесі кезеңде деректер жинау, аннотациялау және модельдерді оқыту жұмыстары жүргізіледі.

Терең оқыту әдістеріне негізделген нейрондық желілер жол бетінде пайда болатын жарықтар мен ойықтарды сенімді түрде сегменттеп, классификациялай алады. Терең оқыту нейрондық желілері жоғары дәлдікке ие, модель көлемі мен есептеу шығындары нақты уақытта жұмыс істеуге шектеу қояды. Осы әдістерді жол бетінің жарықтарын автоматты анықтау жүйелеріне бейімдеу жолдары мен бағалау критерийлері қарастырылады.

Жол бетіндегі пайда болатын жарықтарды анықтау – көлік қозғалысы қауіпсіздігі мен қозғалыстың жайлылығын қамтамасыз етудегі басты мәселелердің бірі. Жарықтар жол бетіндегі кішкентай жіңішке сызықтар түрінде басталып, уақыт өте келе ұлғайып, асфальттың не бетонның беріктігін төмендетеді. Мұндай зақымдануларды ерте анықтап, қажетті жөндеу жұмыстарына кірісу қаржы мен уақытты үнемдеуге, сонымен қатар апаттық жағдайлардың алдын алудың тиімді жолын қарастыруға және жол сапасының тұрақты дамуын қамтамасыз етуге мүмкіндік береді.

Жол бетінің зақымдануын анықтау кезінде терең оқытудың архитектуралары қолданылады. Модельді оқыту және тестілеу үшін Crack500 деректер жиынтығы таңдалынып отыр, себебі ол әртүрлі типтегі және масштабтағы жарықтарды қамтитын жоғары сапалы суреттерді ұсынады. Crack500 — жол бетіндегі жарықтарды анықтауға арналған ең танымал және кеңінен қолданылатын деректер жиынтықтарының бірі. Бұл жиынтық әртүрлі жарық құрылымдарын және әртүрлі түсіру жағдайларын қамтитын жоғары сапалы кескіндерден тұрады. Crack500 деректер жиынтығы әртүрлі масштабтағы және құрылымдағы жарықтарды қамтитындықтан, модельдің сызықтық құрылымдарды дәл анықтау қабілетін арттыруға мүмкіндік береді.

Сонымен қатар, модельдің шынайы ортадағы жұмыс қабілетін бағалау үшін CFD (Crack Forest Dataset) деректер жиынтығында қосымша тестілеу жүргізіледі. Crack Forest Dataset деректер жиынтығы табиғи орта жағдайларында түсірілген жол беттерінің кескіндерін қамтиды және модельдің шынайы ортада жұмыс қабілетін тексеру үшін қосымша дерек көзі ретінде пайдаланылады. Қолданылатын деректер жиынтығының ерекшеліктері мен қолданылу тиімділігі 1-кестеде көрсетілген.

Кесте 1. деректер жиынтығының ерекшеліктері мен қолданылу тиімділігі

Деректер жиынтығы атауы	Ерекшеліктері	Қолдану тиімділігі
Crack500	Жоғары сапалы RGB суреттері, әрқайсысында жарықтар дәл белгіленген. Маскалар толық	DeepCrack++ тәрізді сызықтық құрылымды модельдерге ең жиі қолданылады.

	қолжетімді.	
CFD (Crack Forest Dataset)	Табиғи жағдайдағы әртүрлі жарық түрлері бар және маскалар толық белгіленген.	Модель шынайы ортада жұмыс істейді.

CrackNet, DeepCrack++ архитектураларын оқыту және тестілеу үшін Crack500 және Crack Forest Dataset деректер жиынтығы таңдалынды және заманауи конволюциялық нейрондық желі оқытылады. Жарықтарды пиксель деңгейінде қарастыратын толық немесе жартылай сегментация әдістері әрбір пиксельдің жарыққа тиесілілігін дәл анықтап, зақымданудың нақты формасын қалпына келтіруге мүмкіндік береді. CrackNet, DeepCrack++ архитектуралары дәл осы қағиданы ұстанып, сызықтық құрылымдарды көпмасштабты деңгейде зерттеу, қосымша қабаттарда өңдеу әдістерін пайдаланады. CrackNet архитектурасы жол төсеміндегі жарықтарды анықтау үшін арнайы жасалған және жарықтардың сызықтық құрылымын дәл сипаттауға бағытталған. Бұл модельдің негізгі артықшылығы — құрылымдық ерекшеліктерді сақтай отырып, жарықтарды пиксель деңгейінде анықтай алуы және жол инфрақұрылымын бақылауға бағытталған зерттеулер мен нақты қолданбалы жобаларда сенімді таңдау болып табылады. DeepCrack++ архитектурасы жіңішке, үзілісті жарықтарды да дәл айқындай алады және шағын деректер жиынтығымен жақсы нәтижелер көрсете алады. Модельдің encoder-decoder құрылымы мен контекстік ақпаратты терең өңдеуі оны жол зақымдануларын автоматты түрде, жоғары дәлдікпен анықтауға ең қолайлы архитектуралардың біріне айналдырады.

Бейнелерді өңдеу кезеңінде контрастты күшейту және шуды азайту алгоритмдерін қолдана отырып, жарық пен көлеңке артефактілері барынша жойылды. Эксперимент нәтижелері жүйенің өңдеу жылдамдығы мен бөлектеу дәлдігі арасындағы оңтайлы тепе-теңдікті қамтамасыз ететінін көрсетті.

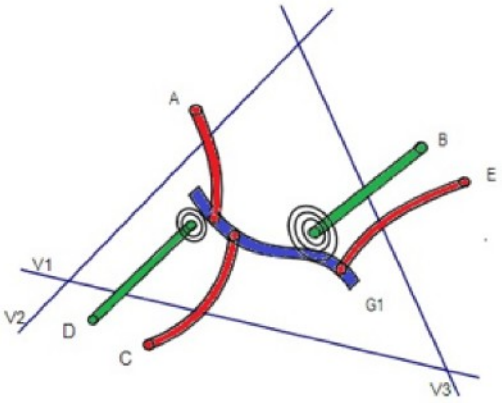
Зерттеу жұмысы жол төсеміндегі жарықтарды, соның ішінде көзге анық байқалмайтын микрожарықтарды анықтауға бағытталған. Осы мақсатта жарықтарды көпөлшемді талдауға мүмкіндік беретін «binding–pointing–crossword» негіздемесі қолданылды. Аталған негіздеме жарықтардың кеңістіктік орналасуын, бағытталуын және өзара қиылысу ерекшеліктерін кешенді түрде сипаттауға мүмкіндік береді.

Binding–pointing–crossword негіздемесі аясында шұңқырлардың қалыптасуы келесі кезеңдер арқылы түсіндіріледі:

1. Pointing кезеңі – жол бетінде микрожарықтардың пайда болуы және олардың локальды бағытталуы (A, C, E).
2. Binding кезеңі – микрожарықтардың өзара байланысып, құрылымдық желі түзуі (B, D).
3. Crossword кезеңі – жарықтардың қиылысуы мен қабаттасуы нәтижесінде кернеу концентрациясының артуы (G1 аймағы).

Терең оқыту әдістері жол төсемінде пайда болатын жарықтарды сенімді түрде сегментациялай отырып, классификациялай алады. Сегментациялау барысында суреттің қай жерінде жарық бар, қай жерінде басқа объектілер орналасқанын анықтайтын маска жасалады. Binding карта көрші орналасқан пиксельдердің арасындағы «binding–pointing–crossword» байланысты бейнелей отырып, жіңішке жарықтарды көлеңкеден, жол белгілерінен ажырата алады. Pointing карта жарықтың траекториясын көрсететін әрбір пиксельдің бағыттаушы векторын береді Нақты шешімдерді қабылдау үшін pointing және crossword модульдерінде қолдануға болады.

1-суретте «binding–pointing–crossword» негіздемесінің схемасында жол төсеміндегі жарықтың орналасу аймағы V_1, V_2, V_3 сызықтарымен шектелген. G_1 жарығына түсірілген A,C,E қисықтары жол төсемінің жарығында орналасқан, B және D қисықтары жол төсеміндегі жарықты талдауда binding профилін бейнелейді.



1-сурет. «Binding–pointing–crossword» негіздемесі

Жарықтарды анықтауда «binding–pointing–crossword» негіздемесін пайдалану көпөлшемді сипаттамаларды бөліп көрсете отырып, аймақтар мен қиылыстарды нақты ажыратуға мүмкіндік берді. Жарықтардың бейнесіндегі көзге көрінбейтін микрожарықтарды анықтау осы негіздеменің көмегімен жүзеге асты. Егер микрожарықтарды уақытында анықтамаса, олардың ішіне әртүрлі лас заттар және су кіріп, жарықтың ұлғаюына алып келеді.

Негізгі аймақтың ішкі маскасында қабаттасудың центрі бола алатын нүкте таңдалынып және G^{mask} өлшеміне тең T^{mask} аймағы қиып алынды. Маскалар аймағында екі бейне ығысуының $\overline{G_c}$ орта мәні есептеледі:

$$\overline{G_c}=\frac{1}{k}\sum_{s} G_c(s) \bullet (1-G^{mask}(s))$$

мұндағы k- пиксель саны, $G_c(s)$) – с каналындағы G жол жарығы суретінің s пиксельдік мәні, $G^{mask}(s)$ – жол жарығының бинарлы маскадағы пиксельдік мәні.

Жол бетінің жарықтарын бейнеге орналастырғанда маска астындағы зақымдану мәндері ғана қарастырылады:

$$M(s) = G^{mask}(s) \cdot T^{mask}(s)$$

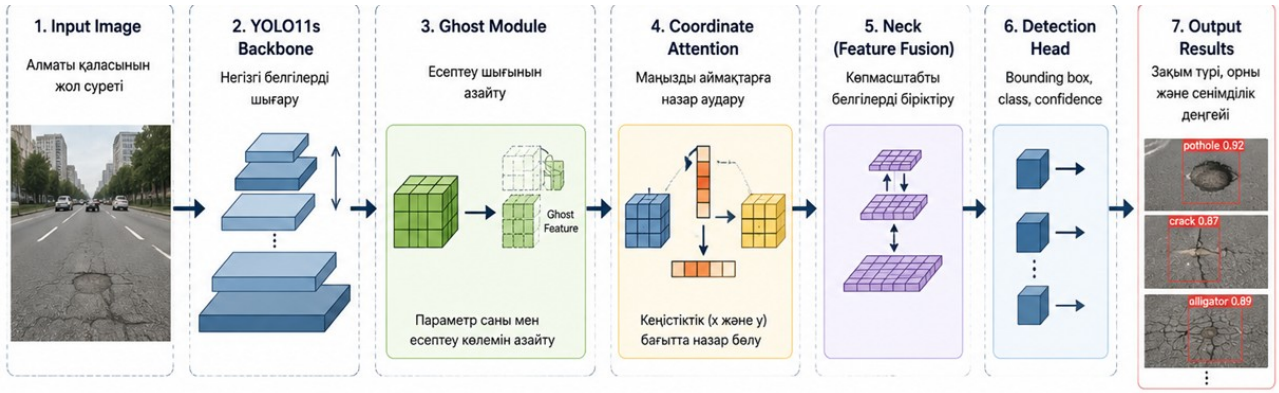
G1 аймағы жарықтардың қиылысу және қабаттасу нүктесін бейнелейді. Диаграммада көрсетілген шеңберлі белгілер осы нүктеде кернеудің жинақталуын көрсетеді. Дәл осы аймақта:

- микрожарықтар макрожарықтарға айналады;
- материалдың тұтастығы бұзылады;
- жарықтардың ені мен тереңдігі артады.

Жол төсеміне ылғал мен ұсақ бөлшектердің енуі жеңілдейді, нәтижесінде төменгі қабаттардың әлсіреуі басталады.

Бейнелердің кеңістіктік қасиеттерін сақтау мақсатында, маска аймағы бейненің орталық бөліктеріне бағытталып таңдалады. Бұл тәсіл жол бетінің зақымдануларын бейнелеу кезінде ақпараттың толық сақталуына мүмкіндік береді және терең оқыту модельдерінің оқыту тиімділігін жоғарылатады, бұл модельдің оқыту кезінде контекстік ақпаратты тиімді пайдалануға мүмкіндік береді. Оқыту таңдамасының ақпараттылығын арттыру үшін бір бейнеде 1-ден 6-ға дейін жол бетінің жарықтары қарастырылды. Нәтижесінде оқыту таңдамасы 3000-ға жуық бейнені құрады. ерттеу барысында Алматы қаласының жолдарынан алынған 3000-ға жуық суреттен тұратын деректер жиыны қалыптастырылды. Деректер жиыны модельді оқыту, валидациялау және тестілеу мақсатында 70:20:10 қатынасында train, validation және test бөліктеріне бөлінді. Train жиыны модельді оқыту үшін, Validation жиыны оқыту барысындағы нәтижені бақылау үшін, ал Test жиыны модельдің соңғы дәлдігін бағалау үшін қолданылды.

2- суретте негізгі архитектура ретінде YOLO11s алынды, себебі ол нақты уақытта объектілерді анықтауға қолайлы және жол зақымдарының орналасуын bounding box арқылы көрсете алады. Модельге Ghost Module енгізу арқылы есептеу шығыны азайтылды, ал Coordinate Attention механизмі зақым орналасқан маңызды аймақтарға назар аударуға мүмкіндік берді. Нәтижесінде модель жол суретінен зақымның түрін, орналасуын және сенімділік деңгейін анықтайды.



Сурет 2– Ghost-CA-YOLO архитектурасы

3-суретте Ghost-CA-YOLO моделінің нәтижелерін бағалау кезеңдері көрсетілген. Алдымен модельге Алматы қаласының жол суреті енгізіледі. Модель суреттегі жол зақымдарын анықтап, олардың түрін, орналасуын және сенімділік деңгейін болжайды. Кейін алынған болжам нәтижелері алдын ала белгіленген нақты деректермен, яғни ground truth белгілерімен салыстырылады. Салыстыру нәтижесінде модельдің дұрыс және қате анықтаулары анықталады. Егер модель зақымды дұрыс тапса, ол True Positive (TP) деп есептеледі. Егер модель зақым жоқ жерде зақым бар деп көрсетсе, ол False Positive (FP) болады. Ал нақты бар зақымды модель таппай қалса, ол False Negative (FN) деп белгіленеді. Осы көрсеткіштер негізінде модельдің сапасы бірнеше метрика арқылы бағаланады. Precision модель анықтаған нысандардың қаншасы дұрыс екенін көрсетеді. Recall нақты бар зақымдардың қаншасын модель тапқанын сипаттайды. F1-score Precision мен Recall арасындағы теңгерімді көрсетеді. mAP@50 және mAP@50–95 модельдің зақымдарды қаншалықты дәл анықтайтынын бағалайды. Сонымен қатар FPS көрсеткіші модельдің нақты уақытта жұмыс істеу жылдамдығын сипаттайды. Confusion Matrix әрбір зақым түрі бойынша модельдің дұрыс және қате жіктеулерін көрсетеді. Ал PR Curve Precision мен Recall арасындағы байланысты талдауға мүмкіндік береді. Осы метрикалардың барлығы модельдің дәлдігін, толықтығын, тұрақтылығын және нақты уақыт режимінде қолдануға жарамдылығын жан-жақты бағалау үшін қолданылады.



Сурет 3 - Бағалау метрикалары

Precision — модель анықтаған зақымдардың қаншасы шынымен дұрыс екенін көрсетеді.

$$Precision = \frac{TP}{TP + FP}$$

Мысалы, модель 100 зақым тапты. Оның 90-ы дұрыс, 10-ы қате болса:

$$Precision = \frac{90}{90 + 10} = 0.90$$

Яғни дәлдік — 90%.

Recall — нақты бар зақымдардың қаншасын модель тапқанын көрсетеді.

$$Recall = \frac{TP}{TP + FN}$$

Мысалы, суреттерде нақты 100 зақым бар. Модель оның 86-сын тапты, 14-ін таппады:

$$Recall = \frac{86}{86 + 14} = 0.86$$

Яғни модель нақты зақымдардың 86% тапты.

F1-score — Precision мен Recall арасындағы теңгерімді көрсетеді.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$Precision = 0.90$$

$$Recall = 0.86$$

$$F1 = 2 \cdot \frac{0.90 \cdot 0.86}{0.90 + 0.86} = 0.88$$

Модель	Precision	Recall	F1-score	mAP@50	FPS
YOLO11s	0.86	0.82	0.84	0.87	42
Ghost-CA-YOLO	0.90	0.86	0.88	0.91	48

Эксперимент барысында ұсынылған Ghost-CA-YOLO моделінің тиімділігі базалық YOLO11s архитектурасымен салыстырылды. Модельдер Алматы қаласының жол суреттері негізінде оқытылып, тестілік деректер арқылы бағаланды.

Бағалау нәтижесінде Ghost-CA-YOLO моделі жол төсеміндегі зақымдарды анықтау кезінде жоғары дәлдік көрсетті. Ghost Module есептеу шығынын азайтуға мүмкіндік берсе, Coordinate Attention механизмі зақым орналасқан маңызды аймақтарды дәлірек анықтауға ықпал етті.

Нәтижесінде модель шұңқыр, жарықшақ, торлы жарықшақ және жамалған асфальт сияқты зақым түрлерін нақты анықтап, олардың орналасуын bounding box арқылы көрсетті. Алынған нәтижелер модельдің жол зақымдарын нақты уақытта анықтауға жарамды екенін дәлелдейді.

Ghost-CA-YOLO моделі базалық YOLO11s моделіне қарағанда зақымдарды дәлірек анықтап, нақты уақыт режимінде тиімді жұмыс көрсетті. Нақты мәндер эксперимент аяқталғаннан кейін енгізіледі.

Қаржылық-технологиялық экожүйелердегі интеллектуалды агенттік шешім қабылдау жүйелері

Ақылаев Жасұлан Ақжолұлы
E-mail-j.akylayev@gmail.com
әл-Фараби атындағы ҚазҰУ

Аңдатпа. Бұл мақала қаржылық-технологиялық экожүйелерде интеллектуалды агенттік шешім қабылдау жүйелерін қолданудың теориялық мүмкіндіктерін қарастыруға арналған. Мақалада интеллектуалды агенттердің қаржылық деректерді өңдеу, тәуекелдерді анықтау, нарықтық өзгерістерді болжау және басқарушылық шешімдерді қолдау үдерісіндегі рөлі сипатталады. Сонымен қатар, интеллектуалды агенттік жүйелерді автоматтандырылған қаржылық басқару үдерістеріне енгізудің ерекшеліктері қарастырылады. Мұндай технологияларды қолданудың шешім қабылдау жеделдігі мен қаржылық басқару тиімділігін арттырудағы маңызы айқындалады.

Түйін сөздер: қаржылық технологиялар, қаржылық экожүйе, интеллектуалды агент, шешім қабылдау, автоматтандырылған басқару.

Кіріспе. Қазіргі кезеңде әлемдік қаржы жүйесінің цифрлық трансформациясы қаржылық-технологиялық экожүйелердің құрылымы мен оларды басқару тәсілдерін түбегейлі өзгертуде. Үлкен деректер, жасанды интеллект, машиналық оқыту, бұлттық есептеулер, блокчейн және автоматтандырылған аналитикалық платформалардың дамуы қаржы ұйымдарына ақпаратты бұрынғыдан әлдеқайда жылдам өңдеуге және басқарушылық шешімдерді деректер негізінде қабылдауға мүмкіндік береді. Сонымен қатар қаржы нарығындағы операциялардың күрделенуі, ақпарат көлемінің үздіксіз өсуі, транзакциялардың нақты уақыт режимінде орындалуы және тәуекел факторларының жылдам өзгеруі дәстүрлі шешім қабылдау тәсілдерінің мүмкіндіктерін шектейді. Осындай жағдайда ақпаратты тек жинақтап немесе талдап қана қоймай, қоршаған ортадағы өзгерістерді анықтайтын, бірнеше баламаны салыстыратын және басқарушылық шешімдерді қолдайтын интеллектуалды агенттік жүйелерді пайдалану ерекше маңызға ие болып отыр.

Қаржылық-технологиялық экожүйе банктерді, сақтандыру ұйымдарын, төлем сервистерін, FinTech компанияларын, инвестициялық платформаларды, мемлекеттік реттеушілерді, кәсіпкерлер мен тұтынушыларды біртұтас цифрлық ортада байланыстыратын күрделі жүйе ретінде қалыптасуда. Мұндай экожүйеде басқарушылық шешімнің сапасы көбінесе әртүрлі дереккөздерден түсетін ақпаратты қаншалықты жылдам өңдеуге, тәуекелдерді ерте анықтауға және ықтимал өзгерістерді болжауға байланысты. Қазіргі зерттеулерде жасанды интеллекттің қаржы саласындағы рөлі жекелеген болжамдық модельдерді пайдаланудан адамның қатысуымен немесе автономды түрде әрекет ететін кешенді шешім қабылдау жүйелеріне қарай ауысып келе жатқаны көрсетіледі. Мәселен, Gang Kou және әріптестері қаржы саласындағы жасанды интеллекттің дамуын ақпаратты алу, болжау, ұсыным қалыптастыру, шешімді бекіту, орындау және мониторинг жүргізу кезеңдерін біріктіретін адам–AI гибридті жүйелеріне өту ретінде қарастырады.

Интеллектуалды агент дәстүрлі бағдарламалық құралдан белгілі бір деңгейдегі автономдылығымен, қойылған мақсатқа сәйкес әрекет етуімен, сыртқы ортадан алынатын ақпаратқа жауап беруімен және жинақталған мәліметтер негізінде өзінің іс-әрекетін бейімдей алуымен ерекшеленеді. Қаржы саласында мұндай агенттер клиенттердің транзакциялық мінез-құлқын талдау, қаржылық тәуекелдерді бағалау, алаяқтық операцияларды анықтау, инвестициялық шешімдерді қолдау, нарықтағы өзгерістерді бақылау, кредиттік тәуекелдерді бағалау және нормативтік талаптардың сақталуын қадағалау сияқты міндеттерді орындауға бейімделе алады. Rizinski мен Trajanov интеллектуалды AI-агенттердің алгоритмдік сауда, алаяқтықты анықтау,

кредиттік тәуекелді бағалау, роботтандырылған қаржылық кеңес беру және RegTech салаларында қолданылатынын атап көрсетеді[1]. Авторлар агенттік тәсілдің маңызды ерекшеліктері ретінде мақсатқа бағытталған мінез-құлықты, үйрену қабілетін және белгілі бір деңгейдегі автономды әрекетті бөліп көрсетеді.

Осы бағыттағы жаңа зерттеулер интеллектуалды агенттерді жеке алгоритм ретінде емес, өзара әрекеттесетін агенттер жүйесі ретінде ұйымдастырудың мүмкіндіктерін көрсетуде. Liu, Li, Nie және Wang FinTech экожүйелерінің ірі көлемді, әртекті және өзара тығыз байланысқан деректер ортасын қалыптастыратынын көрсетіп, түсіндірілетін агенттік жасанды интеллект негізіндегі шешім қабылдау құрылымын ұсынған. Олардың моделінде автономды агенттер мақсатқа бағытталған талдау, бейімделу, бірлескен әрекет және үздіксіз үйрену функцияларымен біріктіріледі. Бұл тәсіл қаржылық шешімнің тек жылдам қабылдануына емес, сонымен бірге оның қалыптасу логикасының түсінікті және бақыланатын болуына бағытталған[2].

Қаржылық шешімдерді қолдау жүйелерінің маңыздылығы туралы Pahsa зерттеуінде де көрсетілген. Автор FinTech саласындағы жасанды интеллект пен машиналық оқыту технологияларын шешімдерді қолдау жүйелерімен біріктіру қаржылық өнімдер мен қызметтерді ұсыну, тәуекелдерді басқару және қызмет көрсету үдерістерінің тиімділігін жоғарылатуға мүмкіндік беретінін атап өтеді. Осы тұрғыдан интеллектуалды агенттік жүйені қаржылық деректерді талдаудың жеке құралы ретінде емес, «деректерді қабылдау – өңдеу – тәуекелді анықтау – сценарийлерді бағалау – балама шешімдерді қалыптастыру – басқарушылық шешімді қолдау» тізбегін қамтамасыз ететін кешенді технологиялық механизм ретінде қарастыру орынды.

Қазақстан үшін де бұл мәселе өзекті. Елдің қаржы секторындағы цифрландыру үдерістері жасанды интеллект пен FinTech шешімдерін енгізуге қолайлы технологиялық ортаны қалыптастыруда. Қазақстандық зерттеушілер Р.Ж. Джусупова, А.С. Джусупов және А.М. Казамбаева жасанды интеллект технологияларын ұлттық қаржы жүйесіне енгізу операциялардың тиімділігін арттыруға, транзакциялық шығындарды төмендетуге және қаржылық үдерістердің ашықтығын қамтамасыз етуге жаңа мүмкіндіктер беретінін атап көрсетеді[3]. Авторлар Қазақстанның қаржы секторындағы цифрлық трансформация жағдайында AI технологияларының маңызының артып келе жатқанын көрсетеді. Сондықтан интеллектуалды агенттік шешім қабылдау жүйелерін отандық қаржылық-технологиялық экожүйелердің ерекшеліктеріне бейімдеу теориялық тұрғыдан ғана емес, қолданбалы жағынан да маңызды.

Сонымен бірге интеллектуалды агенттерді қаржы жүйесіне енгізу бірқатар күрделі мәселелерді туындатады. Біріншіден, қаржылық деректердің дәлдігі, толықтығы және қауіпсіздығы қамтамасыз етілуі қажет. Екіншіден, интеллектуалды агент қалыптастырған шешімдердің түсіндірмелілігі маңызды, өйткені қаржы саласында автоматтандырылған шешімдер клиенттің қаржылық жағдайына, ұйымның тәуекел деңгейіне және жалпы нарықтық тұрақтылыққа ықпал етуі мүмкін. Үшіншіден, агенттердің автономдылық деңгейі мен адамның бақылауы арасындағы теңгерім сақталуы тиіс. 2026 жылғы зерттеулерде автономды AI-агенттердің қаржы нарығындағы мүмкіндіктерімен қатар олардың шешімдерінің ашық еместігі, мақсаттардың сәйкес келмеуі және шамадан тыс орталықтандыру сияқты тәуекелдер де көрсетілген[4].

Осыған байланысты зерттеудің мақсаты – қаржылық-технологиялық экожүйелерде басқарушылық шешімдерді қолдауға арналған интеллектуалды агенттік жүйелердің теориялық негіздерін жүйелеу және оларды қаржылық деректерді өңдеу, тәуекелдерді бағалау, болжам жасау және шешім баламаларын қалыптастыру үдерістерінде қолданудың тиімді бағыттарын анықтау.

Зерттеу мақсатына сәйкес интеллектуалды агенттік жүйелердің қаржы саласындағы мазмұны мен функционалдық ерекшеліктерін анықтау; олардың шешім қабылдауды қолдау жүйелеріндегі орнын негіздеу; қаржылық деректерді интеллектуалды өңдеудің негізгі кезеңдерін жүйелеу; агенттердің тәуекелдерді анықтау

мен болжам жасау мүмкіндіктерін қарастыру; интеллектуалды агенттік жүйенің құрылымдық моделін қалыптастыру және оны қаржылық-технологиялық экожүйелерге енгізудің қолданбалы мүмкіндіктерін бағалау міндеттері белгіленді.

Зерттеуде жүйелік және салыстырмалы талдау, ғылыми әдебиеттерді контенттік талдау, құрылымдық-функционалдық талдау және нәтижелерді салыстырмалы бағалау әдістері қолданылды. Зерттеудің концептуалдық моделі қаржылық деректерді бастапқы жинаудан басқарушылық шешімге дейінгі үдерісті өзара байланысты кезеңдер ретінде қарастырады. Зерттеу нәтижелері интеллектуалды агенттік жүйелерді қаржылық-технологиялық экожүйелерге интеграциялау үлкен көлемдегі қаржылық ақпаратты жедел өңдеуге, тәуекел белгілерін ерте кезеңде анықтауға, ықтимал сценарийлерді салыстыруға және басқарушылық шешімдердің баламаларын автоматты түрде қалыптастыруға мүмкіндік беретінін көрсетті. Зерттеу негізінде «қаржылық деректер → деректерді өңдеу → интеллектуалды агенттер → тәуекелдер мен сценарийлерді бағалау → шешім баламаларын қалыптастыру → басқарушылық шешім» құрылымдық тізбегі ұсынылады. Нәтижелер интеллектуалды агенттің басқарушы тұлғаны толық алмастыруынан гөрі, ақпараттық және аналитикалық жүктемені азайтатын, шешімдердің негізділігін арттыратын және өзгермелі қаржылық ортаға жедел жауап беруге мүмкіндік беретін шешімдерді қолдау құралы ретінде пайдаланылуы неғұрлым орынды екенін көрсетеді.

Зерттеу әдісі мен материалдары. Зерттеу шолу сипатындағы мақала ретінде жүргізілді. Зерттеу материалдары ретінде қаржылық-технологиялық экожүйелер, интеллектуалды агенттік жүйелер, жасанды интеллект және шешім қабылдауды қолдау технологияларына арналған отандық және шетелдік ғылыми еңбектер пайдаланылды. Зерттеуде екі негізгі әдіс қолданылды: әдебиеттерді жүйелі шолу әдісі және салыстырмалы талдау әдісі. Жүйелі шолу арқылы интеллектуалды агенттік жүйелердің қаржы саласында қолданылуы туралы ғылыми зерттеулер жинақталып, негізгі теориялық тұжырымдар жүйеленді. Салыстырмалы талдау әдісі арқылы әртүрлі интеллектуалды шешім қабылдау тәсілдерінің мүмкіндіктері, қолдану бағыттары мен ерекшеліктері салыстырылды. Алынған мәліметтер негізінде қаржылық-технологиялық экожүйелердегі интеллектуалды агенттік жүйелердің қазіргі даму бағыттары қорытындыланды.

Талқылау мен Нәтижелері. Қаржылық-технологиялық экожүйелердің қарқынды дамуы басқарушылық шешімдерді қабылдаудың дәстүрлі тәсілдерін қайта қарастыру қажеттілігін қалыптастырып отыр. Қаржы саласында ақпарат көлемінің үздіксіз артуы, операциялардың нақты уақыт режимінде жүзеге асырылуы және тәуекелдердің жылдам өзгеруі шешім қабылдау үдерісіне интеллектуалды технологияларды енгізудің маңызын арттырады. Осы тұрғыдан алғанда, интеллектуалды агенттік жүйелер қаржылық ақпаратты автоматты түрде жинақтау мен өңдеуді ғана емес, сонымен қатар тәуекелдерді бағалау, ықтимал сценарийлерді салыстыру және басқарушылық шешімдердің бірнеше баламасын ұсыну мүмкіндігін қамтамасыз ететін құрал ретінде қарастырылады.

Ғылыми әдебиеттерді шолу нәтижелері интеллектуалды агенттік жүйелердің негізгі артықшылықтарының бірі олардың автономды және бейімделгіш сипатымен байланысты екенін көрсетеді. Қарапайым автоматтандырылған жүйелер алдын ала белгіленген алгоритмдер бойынша жұмыс істесе, интеллектуалды агенттер өзгермелі ақпараттық ортаға бейімделіп, жаңа деректер негізінде әрекет ету стратегиясын өзгерте алады. Бұл ерекшелік қаржы нарығы үшін маңызды, себебі валюталық бағамдардың, активтер құнының, клиенттік сұраныстың және тәуекел факторларының өзгеруі басқарушылық шешімдерді тұрақты түрде жаңартып отыруды талап етеді.

Интеллектуалды агенттік шешім қабылдау жүйелерінің тағы бір маңызды бағыты – қаржылық тәуекелдерді ерте анықтау. Агенттік жүйелер әртүрлі дереккөздерден алынған ақпаратты өзара байланыстыра отырып, дәстүрлі талдау кезінде байқалмауы мүмкін өзгерістер мен ауытқуларды анықтауға мүмкіндік береді. Бұл кредиттік,

операциялық және нарықтық тәуекелдерді бағалау кезінде шешім қабылдаушы тұлғаға қосымша аналитикалық ақпарат ұсынады. Сонымен қатар интеллектуалды жүйелер бірнеше ықтимал сценарийді қатар талдап, олардың әрқайсысына сәйкес шешім нұсқаларын қалыптастыруға мүмкіндік береді.

Алайда интеллектуалды агенттік жүйелерді енгізуді тек технологиялық тиімділік тұрғысынан бағалау жеткіліксіз. Қаржы саласында қабылданатын шешімдердің сенімділігі деректер сапасына, алгоритмдердің түсіндірмелілігіне, ақпараттық қауіпсіздікке және адамның бақылауына тікелей байланысты. Сондықтан интеллектуалды агент басқарушы маманды толық алмастыратын автономды субъект ретінде емес, оның аналитикалық мүмкіндіктерін кеңейтетін шешімдерді қолдау құралы ретінде қарастырылғаны дұрыс. Әсіресе жоғары тәуекелді қаржылық операцияларда соңғы шешімді адам тарапынан бақылау жауапкершілік пен шешімнің негізділігін қамтамасыз ету үшін маңызды.

Ғылыми еңбектерді салыстырмалы талдау қаржылық технологияларда жасанды интеллекттің жекелеген болжамдық алгоритмдерден кешенді интеллектуалды агенттік шешім қабылдау жүйелеріне қарай дамып келе жатқанын көрсетеді. Қазіргі агенттік жүйелер деректерді қабылдау, талдау, мақсатқа бағытталған пайымдау, стратегия қалыптастыру және бақыланатын әрекеттерді орындау функцияларын біріктіреді. Сонымен қатар түсіндірілетін агенттік AI қаржылық шешімдердің дәлдігімен қатар олардың ашықтығы, бейімделгіштігі және адам тарапынан бақылануы мәселелерін шешуге бағытталған. Ғылыми әдебиеттерді салыстырмалы талдау негізінде алынған нәтижелер 1-кестеде жүйеленді.

1-кесте – Қаржылық-технологиялық экожүйелердегі дәстүрлі және интеллектуалды агенттік шешім қабылдау тәсілдерін салыстыру

Бағалау бағыты	**Дәстүрлі/бастапқы тәсіл**	**Интеллектуалды агенттік тәсіл**	**Әдебиет**
Деректерді өңдеу	Жекелеген AI/ML модельдері мен деректерді талдау	Деректерді қабылдау, пайымдау және әрекетті бір жүйеге біріктіру	Rizinski & Trajanov, 2026; Gong, 2026
Шешім қабылдау	Негізінен адамға тәуелді немесе жеке автоматтандырылған модельдер	Автономды немесе жартылай автономды агенттер арқылы шешімді қолдау	Gong, 2026
Тәуекелдерді басқару	Стандартты DSS және ML құралдары	Үздіксіз мониторинг пен агенттік талдауға негізделген тәсіл	Pahsa, 2024; Rizinski & Trajanov, 2026
Түсіндірмелілік	Black-box модельдерде шектеулі	XAI және визуалды аналитика арқылы түсіндірілетін ұсыныстар	Liu et al., 2026
Бейімделу	Алдын ала анықталған модельдер мен ережелер	Мақсатқа бағытталған пайымдау, бейімделу және үздіксіз үйрену	Liu et al., 2026
Адам–AI өзара әрекеті	AI көбіне аналитикалық құрал ретінде	Адаммен бірлескен шешім қабылдау және бақыланатын автономия	Liu et al., 2026; Gong, 2026және

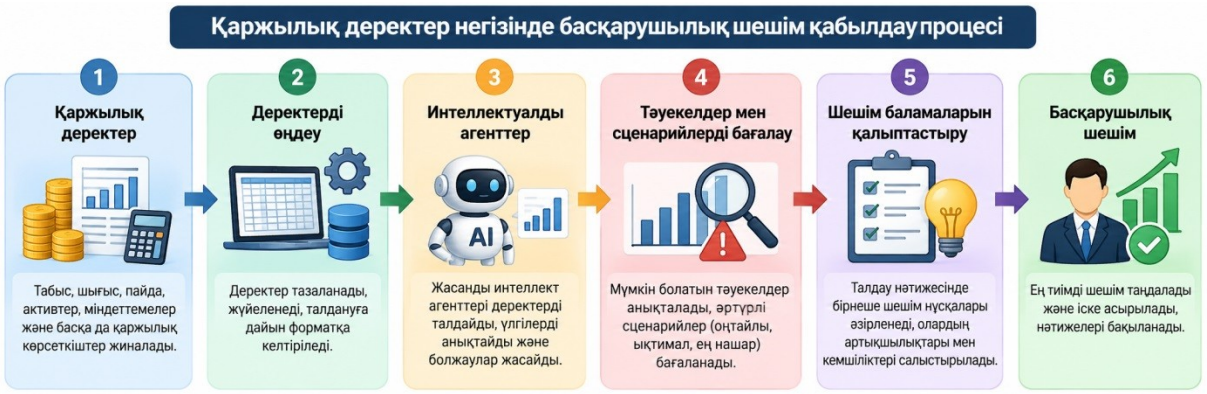
Дереккөз: авторлар Rizinski және Trajanov (2026), Gong (2026), Liu et al. (2026), Pahsa (2024) еңбектерін жүйелі шолу және салыстырмалы талдау негізінде құрастырған.

1-кестеде көрсетілгендей, интеллектуалды агенттік жүйелердің дәстүрлі шешімдерді қолдау технологияларынан негізгі айырмашылығы олардың жеке аналитикалық міндетті орындаумен шектелмеуінде. Агенттік архитектура ақпаратты қабылдаудан бастап пайымдау, стратегия қалыптастыру және бақыланатын әрекетке дейінгі бірнеше кезеңді бір жүйеге біріктіруге мүмкіндік береді. Gong ұсынған модельде қаржылық AI-агент архитектурасы деректерді қабылдау, пайымдау механизмі, стратегия қалыптастыру және бақылаумен орындау сияқты төрт негізгі қабаттан тұрады[5]. Автор сонымен бірге жақын мерзімдегі неғұрлым шынайы қолдану формасы адамның шешім қабылдау үдерісіне енгізілген бақыланатын автономия екенін көрсетеді.

Liu және әріптестерінің зерттеуінде түсіндірілетін агенттік AI тәсілі дәстүрлі black-box модельдерімен салыстырылып, оның шешім дәлдігін, бейімделгіштігін және сенімділігін жақсарту әлеуеті case-driven бағалау арқылы көрсетілген. Ұсынылған жүйе автономды агенттерді түсіндірілетін үлкен деректер аналитикасымен және визуалды аналитикамен байланыстырады. Pahsa еңбегінде AI мен машиналық оқытуға негізделген шешімдерді қолдау жүйелері қаржылық өнімдер, қызмет көрсету арналары және тәуекелдерді басқару процестерін жылдам әрі экономикалық тиімді ұйымдастырудың маңызды бағыты ретінде қарастырылады[6].

Ұсынылған тәсілдің маңызды құрамдас бөлігі – агенттердің өзара әрекеттесу механизмі. Дәстүрлі жүйелерде ақпарат алдын ала анықталған алгоритм бойынша өңделсе, интеллектуалды агенттік ортада жүйе ағымдағы жағдайға байланысты талдау стратегиясын бейімдей алады. Мәселен, күмәнді транзакциялар саны күрт өскен жағдайда тәуекел агентінің басымдығы жоғарылап, қосымша деректерді тексеру процедуралары іске қосылады. Нарықтық көрсеткіштердің елеулі өзгерісі анықталғанда болжамдық агент бірнеше ықтимал сценарийді модельдеп, олардың ықтимал салдарын салыстыруға мүмкіндік береді.

Жүйенің жұмыс логикасын келесі тізбекпен көрсетуге болады:



1-сурет – Қаржылық-технологиялық экожүйеде интеллектуалды агенттер негізінде басқарушылық шешім қабылдау үдерісінің моделі

Ұсынылған модель қаржылық-технологиялық экожүйеде басқарушылық шешім қабылдаудың бірізді және өзара байланысты кезеңдерін көрсетеді. Қаржылық деректерді жинаудан басталған үдеріс оларды өңдеу, интеллектуалды агенттер арқылы талдау, тәуекелдер мен ықтимал сценарийлерді бағалау және шешім баламаларын қалыптастыру кезеңдері арқылы жүзеге асырылады. Нәтижесінде басқарушыға бірнеше ықтимал нұсқаның ішінен негізделген шешім қабылдауға мүмкіндік беріледі.

Интеллектуалды агенттерді қолданудың негізгі артықшылығы – үлкен көлемдегі қаржылық ақпаратты жедел өңдеу, деректер арасындағы жасырын заңдылықтарды анықтау және ықтимал тәуекелдерді алдын ала бағалау мүмкіндігі. Мұндай тәсіл

басқарушылық шешімдердің тек ағымдағы көрсеткіштерге емес, сонымен қатар болжамдық сценарийлерге негізделуіне жағдай жасайды. Сонымен қатар шешім баламаларын салыстыру қаржылық шығындарды, тәуекел деңгейін және күтілетін нәтижелерді кешенді бағалауға мүмкіндік береді.

Осылайша, «Қаржылық деректер → Деректерді өңдеу → Интеллектуалды агенттер → Тәуекелдер мен сценарийлерді бағалау → Шешім баламаларын қалыптастыру → Басқарушылық шешім» тізбегі қаржылық-технологиялық экожүйелердегі шешімдерді қолдаудың тұтас интеллектуалды механизмін сипаттайды.

Қорытынды. Жүргізілген шолу және салыстырмалы талдау қаржылық-технологиялық экожүйелердегі интеллектуалды агенттік шешім қабылдау жүйелерінің қазіргі цифрлық қаржы саласындағы маңызын айқындауға мүмкіндік берді. Қаржылық операциялар көлемінің ұлғаюы, деректер ағынының күрделенуі және нарық жағдайының жылдам өзгеруі ақпаратты жедел өңдеп, ықтимал тәуекелдерді анықтайтын және басқарушылық шешімдерді қолдайтын интеллектуалды технологияларға қажеттілікті арттырып отыр. Зерттеу нәтижелері интеллектуалды агенттік жүйелердің дәстүрлі автоматтандырылған және шешімдерді қолдау жүйелерінен функционалдық мүмкіндіктерінің кеңдігімен ерекшеленетінін көрсетті. Интеллектуалды агенттер қаржылық деректерді жинау мен өңдеумен қатар, ақпаратты талдау, тәуекел факторларын анықтау, ықтимал сценарийлерді бағалау және шешімдердің балама нұсқаларын қалыптастыру міндеттерін біртұтас үдеріс шеңберінде жүзеге асыруға мүмкіндік береді. Әдебиеттерді салыстырмалы талдау агенттік тәсілдердің бейімделу, автономды әрекет ету және адаммен бірлескен шешім қабылдау мүмкіндіктеріне ерекше мән берілетінін көрсетті.

Пайдаланылған әдебиеттер

1. Rizinski M., Trajanov D. AI Agents in Finance and Fintech: A Scientific Review of Agent-Based Systems, Applications, and Future Horizons // Computers, Materials & Continua. – 2026. – Vol. 86, № 1. – P. 1–34. – DOI: https://doi.org/10.32604/cmc.2025.069678
2. Liu H., Li J., Nie Y., Wang Y. Big Data-Driven Explainable Agentic AI Decision Frameworks for Enterprise Innovation in FinTech Ecosystems // Big Data. – 2026. – Vol. 14, № 4. – P. 259–267. – DOI: https://doi.org/10.1177/2167647X261432592
3. Джусупова Р. Ж., Джусупов А. С., Казамбаева А. М. Роль искусственного интеллекта в развитии финансовой системы в Казахстане // Вестник Актюбинского регионального университета имени К. Жубанова. – 2026. – Т. 82, № 4. – DOI: https://doi.org/10.70239/arsu.2025.t82.n4.26
4. Kou G., Li Y., Wang H. et al. Human–AI hybrid finance: from AI tools to decision systems // Financial Innovation. – 2026. – Vol. 12. – Art. 125
5. Gong H. AI Agents in Financial Markets: Architecture, Applications, and Systemic Implications // FinTech. – 2026. – Vol. 5, № 2. – Art. 34. – DOI: https://doi.org/10.3390/fintech5020034
6. Pahsa A. Financial Technology Decision Support Systems // Journal of Electrical Systems and Information Technology. – 2024. – Vol. 11. – Art. 5. – DOI: https://doi.org/10.1186/s43067-023-00130-0

ГИБРИДНАЯ ИНТЕЛЛЕКТУАЛЬНАЯ МОДЕЛЬ ФОРМАЛЬНОЙ ПРОВЕРКИ ЗАЯВОК НА НАУЧНЫЕ И НАУЧНО-ТЕХНИЧЕСКИЕ ПРОЕКТЫ

Мутанов Г.М. [1], Усатова О.А. [1,2], Науменко В.В. [1,3], Карюкин В.И. [1,4], Төреқұл С.М. [1,4], Аманжолова С.Т. [1,5]
e-mail: naumenko.v5@mail.ru

Институт информационных и вычислительных технологий КН МНВО РК, г. Алматы, Казахстан

Алматинский университет энергетики и связи имени Гумарбека Даукеева, г. Алматы, Казахстан

Университет Туран, г. Алматы, Казахстан

Казахский Национальный Университет имени аль-Фараби, г. Алматы, Казахстан

Astana IT University, г. Астана, Казахстан

Аннотация. *Предложена гибридная модель предварительной проверки заявок, объединяющая детерминированные правила, семантический анализ и доказательный протокол. Модель проверяет комплектность, форматы, бюджет, TRL и согласованность цели с результатами, после чего направляет заявку на возврат, ручную проверку либо допуск. Контролируемый эксперимент проведён на 360 заявках, сформированных по фактической схеме Django-прототипа. Гибридный режим достиг полноты 96,8 % и F1 = 91,8 % против 45,5 % и 62,5 % для одних правил; различие статистически значимо по точному критерию Мак-Немара ($p = 3{,}71 \cdot 10^{-12}$). Семантический сигнал используется только для маршрутизации к человеку, что исключает автоматический отказ по вероятностному выводу.*

Ключевые слова: *научная экспертиза, формальная проверка, искусственный интеллект, семантический анализ, гибридная модель, человек в контуре, объяснимость, Django*

1. Введение

Формальная проверка является входным контуром государственной научно-технической экспертизы: до содержательной оценки необходимо установить комплектность заявки, корректность реквизитов, соблюдение ограничений конкурса и внутреннюю согласованность представленных данных. Действующие в Республике Казахстан правила организации ГНТЭ и финансирования научной деятельности задают изменяемый нормативный контекст, поэтому проверка должна быть не набором жёстко встроенных условий, а версионируемой системой требований с фиксируемыми основаниями решения [1; 2]. Международная методология также требует различать фундаментальные исследования, прикладные исследования и экспериментальные разработки, что влияет на допустимые формы результата и набор проверяемых доказательств [3].

Полностью ручной контроль плохо масштабируется и переносит значительную часть административной работы на эксперта. Полностью вероятностная модель, напротив, не гарантирует соблюдение однозначных ограничений и может сформировать правдоподобное, но неверное объяснение. В системах с высокой ценой ошибки

требуется сочетать воспроизводимые правила, статистическую обработку текста и обязательный человеческий контроль неопределённых случаев. Такой подход соответствует принципам тестирования, верификации и валидации ИИ, а также явного распределения ответственности между алгоритмом и оператором [4; 5].

Цель исследования – разработать и экспериментально проверить гибридную модель формального контроля заявок на научные и научно-технические проекты. Научная гипотеза состоит в том, что последовательное объединение детерминированного и семантического контуров повышает полноту обнаружения нарушений без передачи вероятностной модели права на автоматический отказ. Новизна решения заключается в доказательной маршрутизации: каждый вывод связывается с кодом требования, фрагментом заявки, значением показателя и одним из трёх действий – возврат, ручная проверка или допуск.

2. Материалы и методы

Модель разработана в рамках проекта BR28713385 и согласована с архитектурой двух прототипов системы экспертизы. Реализованный Django-контур хранит около 150 полей анкеты, многоязычные аннотации и ожидаемые результаты, повторяющиеся сведения о команде, оборудовании, услугах и командировках, а также восемь файловых слотов [12]. Серверный валидатор проверяет обязательные поля, числовые значения, допустимые перечисления, год, уровни TRL, этап НИОКР, структуру повторителей, равенство общей суммы бюджету по годам, наличие обязательных документов, размер и длину имени файла. Тем самым однозначные нарушения выявляются без обращения к генеративной модели.

Семантический контур предназначен для требований, которые нельзя свести к типу поля или арифметическому равенству: соответствие цели ожидаемому результату, наличие измеримого выхода и признаки логического противоречия. Для промышленной версии предусмотрены локальные языковые модели и извлечение доказательств из конкурсной документации. В эксперименте использован воспроизводимый нижний базовый уровень: символьные TF–IDF n-граммы длиной 3–5 и косинусная близость текстов цели и результата. Такой способ устойчив к части русских словоизменений; в дальнейшем он может быть заменён предложенческими эмбеддингами [6], RAG-контуром с указанием источника требования [7] и моделями анализа визуально насыщенных документов [8].



Рис. 1 – Архитектура гибридной формальной проверки

Для заявки x семантический риск S(x) определяется по близости векторов цели и результата, наличию измеримого показателя M, достаточной информативности L и индикатору отрицания N. Коэффициенты заданы до эксперимента как инженерный приоритет смыслового соответствия; значение ограничивается интервалом [0; 1].

$$S(x) = min[1; 0{,}62(1 - cos(vg, vr)) + 0{,}23(1 - M) + 0{,}15(1 - L) + 0{,}18N] \quad (1)$$

Порог ручной проверки принят равным τ = 0,58. Если срабатывает блокирующее правило, заявка возвращается на доработку с точным описанием нарушения. При отсутствии такого правила и S(x) ≥ τ материал направляется оператору; автоматический отказ по семантическому сигналу запрещён. Если оба контура не выявили риска, заявка допускается к следующему этапу. Такое разделение сохраняет объяснимость решений [9] и уменьшает технический долг, который возникает при скрытом смешении правил, данных и моделей [10].

3. Дизайн эксперимента

Проверка проведена на контролируемом корпусе из 360 обезличенных заявок, автоматически сформированных по реальной схеме прототипа при фиксированном seed 28713385. Корпус включал 140 корректных записей и 220 записей с заранее известными нарушениями: 70 формальных, 60 случаев несоответствия цели и результата, 40 неизмеримых результатов, 30 сочетаний формальных и семантических ошибок и 20 скрытых противоречий. Формальные дефекты включали недопустимый язык и год, обратную динамику TRL, неверный этап НИОКР, нечисловой бюджет, расхождение сумм, пропуск обязательного поля или документа, нарушение структуры повторителя и превышение размера файла. Операционные конкурсные данные не использовались, поэтому эксперимент оценивает способность модели обнаруживать контролируемые дефекты, но не их реальную распространённость.

Сравнивались три режима: только правила, только семантический анализ и их объединение. Рассчитывались precision, recall, F1 и accuracy; для precision и recall построены 95 %-ные интервалы Уилсона [11], для F1 — перцентильный bootstrap по 4000 выборкам. Различие правил и гибридной модели оценивалось точным критерием Мак-Немара на парных результатах. Дополнительно выполнено 12 регрессионных проверок реализованного валидатора; все 12 завершились успешно.

4. Экспериментальные результаты

Экспериментальные результаты представлены в таблице 1.

Таблица 1. Результаты контролируемого эксперимента

Режим	TP	FP	Precision, %	Recall, %	F1, %
Правила	100	0	100.0	45.5	62.5
Семантика	143	31	82.2	65.0	72.6
Гибрид	213	31	87.3	96.8	91.8

Детерминированный режим не сформировал ложных срабатываний, однако обнаружил только 45,5 % всех некорректных заявок: его высокая precision отражает надёжность однозначных правил, а recall = 45,5 % – неспособность видеть смысловые дефекты. Семантический режим повысил recall до 65,0 %, но направил на проверку 31 корректную заявку. Гибридная модель выявила 213 из 220 нарушений: precision составила 87,3% (95% ДИ 82,5 – 90,9), recall – 96,8 % (93,6 – 98,5), F1 – 91,8 % (bootstrap 95% ДИ 89,1 – 94,3), accuracy – 89,4 %. Следовательно, семантический контур

целесообразно трактовать как механизм безопасной эскалации, а не как автоматический классификатор отказа. Также результаты эксперимента представлены в виде гистограмм на рис. 2 и рис. 3.

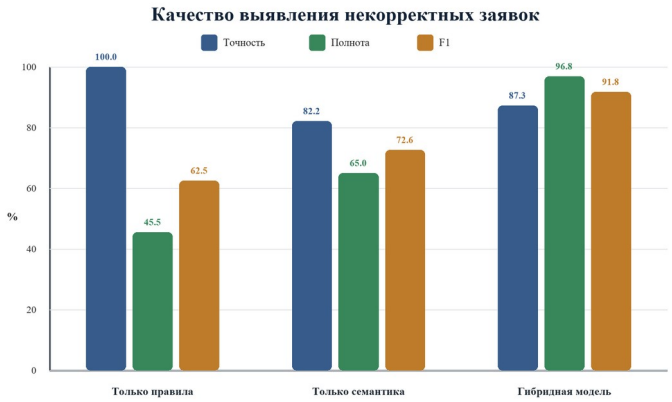


Рис. 2 – Сравнение точности, полноты и F1 трёх режимов

Парное сравнение показало 113 случаев, в которых гибрид был прав, а правила ошибались, и 31 обратный случай; точный тест Мак-Немара дал $p = 3{,}71 \cdot 10^{-12}$. Улучшение не объясняется случайным перераспределением ошибок. По классам гибрид обнаружил 100 % формальных, 100 % неизмеримых и 100 % комбинированных нарушений, 96,7 % несоответствий цели и результата и 75 % скрытых противоречий. Последний показатель определяет основную область дальнейшего развития модели.

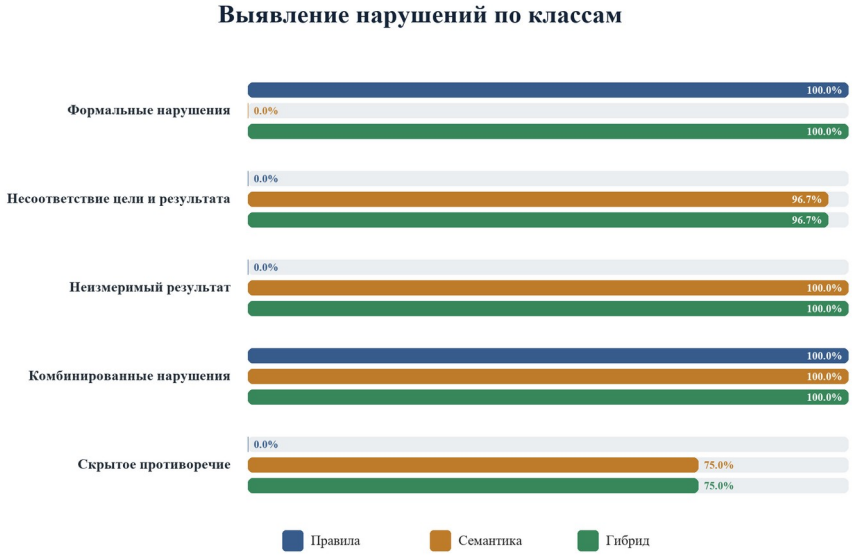


Рис. 3 – Доля обнаруженных нарушений по классам

5. Обсуждение

Результаты подтверждают исходную гипотезу: правила и семантическая модель устраняют разные классы ошибок. Правила гарантируют воспроизводимость там, где требование можно выразить предикатом, диапазоном или арифметическим соотношением. Семантический слой обнаруживает нарушения, для которых важен контекст, но создаёт дополнительную нагрузку ручной проверки. Поэтому 31 ложное срабатывание не является 31 автоматическим отказом: это 31 безопасно эскалированная заявка. Управляемый компромисс задаётся порогом τ и может настраиваться под стоимость пропуска и стоимость дополнительного просмотра.

Доказательный протокол должен хранить версию требования, идентификатор правила, исходное значение, найденный фрагмент, оценку риска и итоговое действие.

Такая структура позволяет воспроизвести проверку после изменения конкурсной документации и отделяет машинный сигнал от административного решения. Для конфиденциальных заявок семантический сервис следует развёртывать локально, а в языковую модель передавать минимально необходимый деперсонализированный контекст. Перед промышленным внедрением требуется независимая разметка реальных обезличенных заявок, междисциплинарная стратификация выборки, оценка калибровки уверенности и мониторинг дрейфа требований.

Ограничения исследования определяются контролируемым характером корпуса. Полученные метрики нельзя интерпретировать как эксплуатационную точность на всех конкурсах. Семантический слой является лёгкой TF–IDF-моделью, а не промышленной LLM; время её работы не включает OCR, базы данных, внешние реестры и инференс большой языковой модели. Вместе с тем такой дизайн полезен как нижняя воспроизводимая граница: улучшенная модель должна сравниваться с тем же корпусом и не должна ухудшать безопасность маршрутизации.

6. Заключение

Разработана гибридная интеллектуальная модель формальной проверки проектных заявок, в которой детерминированные правила отвечают за однозначные требования, семантический контур — за контекстные несоответствия, а человек принимает решение во всех вероятностных случаях. На контролируемом корпусе гибрид повысил F1 с 62,5 до 91,8 % и recall с 45,5 до 96,8 % относительно одних правил. Статистическая значимость подтверждена критерием Мак-Немара. Практический результат состоит не в автоматизации отказа, а в раннем обнаружении дефектов, доказательной обратной связи заявителю и снижении технической нагрузки на последующие этапы научной экспертизы.

7. Благодарность

Данное исследование финансировалось Комитетом науки МНВО РК (программно-целевое финансирование на 2025-2027 годы ИРН BR28713385 «Разработка интеллектуальной системы экспертизы и оценки научных, научно-технических проектов»)

Список литературы:

1. Министерство науки и высшего образования Республики Казахстан. Об утверждении Правил организации и проведения государственной научно-технической экспертизы: приказ от 7 ноября 2024 года № 517, с изменениями. – URL: https://adilet.zan.kz/rus/docs/V2400035344 (дата обращения: 19.09.2026).

2. Министерство науки и высшего образования Республики Казахстан. Об утверждении Правил базового и программно-целевого финансирования научной и (или) научно-технической деятельности: приказ от 6 ноября 2023 года № 563, с изменениями. – URL: https://adilet.zan.kz/rus/docs/V2300033613 (дата обращения: 19.09.2026).

3. OECD. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. – Paris: OECD Publishing, 2015. – 400 p. – DOI: https://doi.org/10.1787/9789264239012-en (дата обращения: 19.09.2026).

4. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). – Gaithersburg: National Institute of Standards and Technology, 2023. – DOI: https://doi.org/10.6028/NIST.AI.100-1 (дата обращения: 19.09.2026).

5. Amershi S., Weld D., Vorvoreanu M. et al. Guidelines for Human-AI Interaction // Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. – 2019. – P. 1–13. – DOI: https://doi.org/10.1145/3290605.3300233.

6. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – P. 3982–3992. – DOI: https://doi.org/10.18653/v1/D19-1410.

7. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459–9474.

8. Peng Q., Pan Y., Wang W. et al. ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding // Findings of the Association for Computational Linguistics: EMNLP 2022. – 2022. – P. 3744–3756. – DOI: https://doi.org/10.18653/v1/2022.findings-emnlp.274.

9. Ribeiro M. T., Singh S., Guestrin C. Why Should I Trust You? Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 1135–1144. – DOI: https://doi.org/10.1145/2939672.2939778.

10. Sculley D., Holt G., Golovin D. et al. Hidden Technical Debt in Machine Learning Systems // Advances in Neural Information Processing Systems. – 2015. – Vol. 28. – P. 2503–2511.

11. Wilson E. B. Probable Inference, the Law of Succession, and Statistical Inference // Journal of the American Statistical Association. – 1927. – Vol. 22, No. 158. – P. 209–212. – DOI: https://doi.org/10.1080/01621459.1927.10502953.

12. Материалы проекта BR28713385. Техническое задание на разработку системы автоматизированной проверки проектных заявок на основе LLM. – Алматы, 2026.

ПИЛОТНАЯ АПРОБАЦИЯ МОДЕЛИ КЛАССИФИКАЦИИ ЭТАПОВ НИОКР НА ОСНОВЕ ФОРМАЛИЗОВАННЫХ ПРАВИЛ

[1,2]Торекул С.М., [1]Мутанов Г.М.

e-mail: saule.torekul@gmail.com

[1]Институт информационных и вычислительных технологий КН МОН РК, Казахстан

[2]Казахский Национальный университет имени аль-Фараби, Казахстан

Аннотация

Предложена объяснимая модель классификации конкурсных заявок на основе формализованных правил по трём этапам НИОКР: фундаментальное, прикладное исследование и опытно-конструкторская работа. Модель включает механизм категориальной оценки уверенности и сопоставляется с базовыми моделями TF-IDF + Logistic Regression и TF-IDF + Linear SVM. Пилотная апробация выполнена на сбалансированной выборке из 108 обезличенных заявок: 72 заявки использованы для настройки правил, 36 – для контрольной оценки. На контрольной выборке модель достигла точности 0,528 и Macro-F1 0,548, тогда как лучшая базовая модель показала 0,694 и 0,682 соответственно. Результаты рассматриваются как пилотное подтверждение реализуемости объяснимой процедуры поддержки эксперта.

Введение

Научная и научно-техническая экспертиза проводится в условиях большого объёма конкурсной документации и роста числа междисциплинарных проектов. При этом экспертные решения могут различаться, зависеть от состава рецензентов и содержать элементы субъективности [1-3]. Низкая согласованность экспертов при оценке одних и тех же заявок также подтверждена экспериментально [4]. Поскольку результаты экспертизы влияют на распределение государственного финансирования, оценка должна быть обоснованной, объективной, прозрачной и воспроизводимой.

В Республике Казахстан государственная научно-техническая экспертиза включает проверку конкурсной документации, независимую экспертизу, обработку результатов и рассмотрение материалов Национальными научными советами [5]. Одним из направлений её развития является цифровизация. В исследованиях предложены архитектура системы оценки научных проектов на основе искусственного интеллекта и блокчейна [6] и рекомендации по созданию интегрированной цифровой платформы для Казахстана [7]. На государственном уровне также отмечается необходимость повышения качества экспертизы, цифровизации научной сферы, прозрачного использования средств и применения искусственного интеллекта [8, 9].

Важной задачей экспертизы является определение фактического этапа НИОКР. Заявленный тип исследования не всегда соответствует содержанию проекта: цели и ожидаемые результаты могут сочетать признаки нескольких этапов или не подтверждаться необходимыми материалами. Поэтому классификация должна учитывать задачи, методы, практическую направленность, ожидаемые результаты и представленные доказательства. Ошибка в определении этапа может привести к применению неподходящих критериев оценки.

Развитие научного результата в общем виде можно представить как переход от фундаментального исследования к прикладному исследованию, а затем к опытно-конструкторской работе. Такое разграничение предусмотрено законодательством Республики Казахстан [10] и согласуется с Руководством Фраскати ОЭСР [11]. Фундаментальные исследования направлены на получение новых знаний и выявление закономерностей; прикладные – на решение практической задачи и проверку применимости научных подходов; ОКР – на разработку и экспериментальную проверку технического, технологического или программного решения. При этом

последовательность «ФИ → ПИ → ОКР» не является строго линейной: проект может включать повторные исследования и признаки нескольких этапов [12].

Следует различать тип конкурса, тип исследования, этап НИОКР и уровень технологической готовности TRL. Тип конкурса определяет механизм финансирования и условия участия, тип исследования является заявленной характеристикой проекта, а этап НИОКР отражает его фактическое содержание и результаты. TRL характеризует готовность конкретной технологии и не является прямым аналогом этапа НИОКР [13, 14]. Поскольку названия и требования конкурсов могут меняться, универсальный алгоритм целесообразно строить на содержательных и доказательных признаках этапов НИОКР. Это позволит применять методику в разных программах финансирования.

Методы обработки естественного языка и машинного обучения применяются для классификации научных текстов, подбора экспертов и предварительного анализа заявок [15, 16]. Однако для государственной экспертизы важна не только точность алгоритма, но и объяснимость результата [17, 18]. Для этого необходимо формализовать обязательные, дополнительные и противоречащие признаки этапов НИОКР, текстовые индикаторы, доказательные результаты и правила принятия решения. Отдельные слова не должны служить самостоятельным доказательством: их необходимо анализировать вместе с целями, задачами и ожидаемыми результатами заявки.

Научная проблема состоит в отсутствии однозначного соответствия между текстом заявки и характеристиками этапов НИОКР. Исследовательский пробел связан с недостаточной разработанностью объяснимых формализованных правил, применимых независимо от типа конкурса. Существующие методы автоматизации чаще направлены на скрининг документов, подбор экспертов и прогнозирование оценок [16-19], тогда как содержательная верификация этапа НИОКР требует отдельного решения.

Цель исследования – разработать объяснимую информационно-аналитическую модель классификации этапов НИОКР на основе формализованных правил, реализовать механизм оценки уверенности и сопоставить результаты с базовыми NLP-моделями на фиксированной тестовой выборке. Вклад работы состоит в формальном описании критериев трёх этапов НИОКР, правилах их применения к многоязычным текстам заявок и выделении случаев, требующих экспертной проверки. Модель предназначена для поддержки эксперта, а не для замены государственной научно-технической экспертизы.

В работе рассматриваются три класса: ФИ – фундаментальное исследование, ПИ – прикладное исследование и ОКР – опытно-конструкторская работа. Исходной референтной меткой служит поле research_type из информационной системы. Поскольку эта метка не подтверждена независимой повторной экспертной разметкой, результаты отражают согласованность модели с имеющейся классификацией, но не абсолютную истинность класса.

Материалы и методы

Источником данных послужила информационная база конкурсных заявок на научные и научно-технические проекты за 2014-2025 гг. Исходная выгрузка содержала 354 802 строки и 25 признаков, относящихся к 67 708 уникальным заявкам (таб.1). Повторения были связаны с соединением основной таблицы с зависимыми сущностями, прежде всего с таблицей вложений.

Таблица 1. Этапы формирования исследовательского датасета

Характеристика	**Значение**
Период наблюдения	2014–2025 гг.
Строк в исходной выгрузке	354 802
Уникальных заявок после дедупликации	67 708
Удалено повторных строк	287 094
Исключено по обязательным требованиям	36 770
Заявок в аналитическом датасете v1	30 938

Признаков в датасете v1	38
Заявок с конфликтующими значениями	10

Предварительная обработка включала нормализацию пробелов и пустых значений, удаление полностью идентичных строк, проверку повторов по application_id и сведение данных к одной записи на заявку. После дедупликации были очищены текстовые поля, унифицированы однозначные варианты research_type и сформированы объединённые тексты на русском, казахском и английском языках.

Аналитический набор данных включает 18 040 прикладных, 12 862 фундаментальных и 36 опытно-конструкторских заявок. Для пилотной проверки был применён дополнительный фильтр полноты: заявка должна была содержать research_type, current_status, название проекта, цель, задачи, ожидаемые результаты и аннотацию на трёх языках. После этого были доступны 17 149 ПИ, 12 665 ФИ и 36 ОКР. Из-за малого числа ОКР сформирована сбалансированная выборка из 108 обезличенных заявок – по 36 документов каждого класса. Отбор и разбиение выполнялись с random_state=42. Выборка для настройки правил включала 72 заявки, тестова выборка - 36 заявок. Обезличивание выполнялось на уровне состава признаков: в экспериментальный набор данных не включались сведения о заявителях, экспертах, контактных данных и персональных идентификаторах. application_id сохранялся только как технический псевдоидентификатор для контроля уникальности и не использовался как модельный признак.

Каждый этап НИОКР представлен как совокупность смысловых признаков и проверяемых доказательств (таб.2). Ключевые слова используются только как индикаторы: сами по себе они не должны определять класс без учёта цели, ожидаемого результата и контекста заявки.

Таблица 2. Формальные критерии классификации этапов НИОКР

Этап НИОКР	**Содержательные признаки**	**Возможные доказательства в тексте заявки**	**Правило решения**
Фундаментальное исследование	Получение новых знаний, выявление закономерностей, объяснение причин и механизмов явлений.	Гипотеза, теория, концепция, научная модель, публикации, формулировки «исследовать», «изучить», «выявить».	Класс выбирается при преобладании познавательной цели и отсутствии признаков готового практического решения или внедрения.
Прикладное исследование	Решение конкретной практической задачи на основе научного подхода.	Методика, алгоритм, модель, программный модуль, рекомендации, лабораторная апробация, оценка применимости.	Класс выбирается при наличии практической проблемы и научно обоснованного способа её решения без завершённой технологии внедрения.
Опытно-конструкторская работа	Реализация научно-технического результата в виде изделия, технологии, системы или программного продукта.	Опытный образец, технические параметры, испытания, конструкторская или эксплуатационная документация, план внедрения.	Класс выбирается при наличии объекта разработки, технической верификации и признаков подготовки к эксплуатации или внедрению.

Общая последовательность работы модели представлена на рисунке 1. Модель анализирует цель, задачи, ожидаемые результаты, аннотацию и название проекта на русском, казахском и английском языках.



Рисунок 1. Общая последовательность работы модели

Для каждого класса с рассчитывается итоговая оценка:

$$S_c = 3M_c + A_c - 3E_c - 2C_c, \quad (1)$$

где M_c – количество обязательных индикаторов класса, A_c – дополнительные индикаторы, E_c – исключающие признаки, C_c – противоречащие признаки. Класс с наибольшей оценкой рассматривается как предварительное решение модели.

Дополнительно рассчитывается разрыв между первой и второй оценкой:

$$\delta = S_{(1)} - S_{(2)}, \quad (2)$$

В работе используется порог δ = 1: если разность между первой и второй оценкой меньше 1, заявка относится к пограничным случаям, поскольку различие между двумя классами недостаточно выражено. Порог δ = 1 был выбран и проверен только на выборке для настройки правил; после просмотра результатов контрольной оценки он не изменялся. Уровень уверенности не интерпретируется как вероятность, поскольку отдельная калибровка модели не проводилась. Он используется как категориальная оценка надёжности предварительного решения, представленная в таблице 3.

Таблица 3. Шкала интерпретации уровня уверенности модели

Уровень уверенности	**Условия**	**Интерпретация**
Высокий	Выполнены минимальные условия выбранного класса, разность оценок ≥ 3, нет существенных противоречий и языкового расхождения.	Решение можно использовать как устойчивую предварительную рекомендацию эксперту.
Средний	Выполнены минимальные условия выбранного класса, 1 ≤ разность оценок < 3, противоречия не являются существенными.	Решение допустимо как предварительное, но желательно экспертное подтверждение.
Низкий	Разность оценок < 1, не выполнены минимальные условия класса, выявлены противоречивые признаки, пограничный случай, неполные данные или расхождение между языковыми версиями.	Заявка требует экспертной проверки.

Для сопоставления с подходом на основе формализованных правил использованы две базовые модели текстовой классификации: TF-IDF + Logistic Regression и TF-IDF + Linear SVM. Они обучались на той же development-части и оценивались на той же фиксированной test-части. Из модельных признаков исключались research_type, research_type_original, current_status, overall_type, competition_type и служебные поля классификации на основе формализованных правил.

Для внутренней проверки на выборке для настройки правил применялась StratifiedKFold(n_splits=4, shuffle=True, random_state=42). Основными метриками

выбраны balanced accuracy и Macro-F1, поскольку они одинаково учитывают все три класса.

Результаты

Разведочный анализ данных (EDA) показал сильный дисбаланс классов: в аналитическом датасете первой версии содержится 18 040 прикладных заявок, 12 862 фундаментальных и только 36 ОКР. Средняя длина объединённого текста составила 2130,87 слова для ПИ, 2151,73 слова для ФИ и 2394,94 слова для ОКР. Однако показатель для ОКР менее устойчив из-за малого числа объектов.

На контрольной выборке принудительная классификация на основе формализованных правил, при которой для каждой заявки выбирается класс с максимальной оценкой, показала сбалансированную точность (balanced accuracy) = 0,528 и Macro-F1 = 0,548. По отдельным классам F1 составила: ФИ = 0,632, ПИ = 0,467, ОКР = 0,545. В селективном режиме модель приняла окончательное решение по 14 из 36 заявок контрольной выборки, то есть при покрытии 38,9%; на подмножестве заявок с принятым решением сбалансированная точность составила 0,400, Macro-F1 - 0,322. Итоговое сравнение модели на основе формализованных правил и базовых сравнительных моделей на одной контрольной выборке приведено в таблице 4.

Таблица 4. Сравнение модели на основе формализованных правил и базовых сравнительных моделей на контрольной выборке

Модель	**Сбалансированная точность**	**Macro-F1**	**F1 ФИ**	**F1 ПИ**	**F1 ОКР**	**Покрытие**
Формализованные правила, принудительный режим	0,528	0,548	0,632	0,467	0,545	1,000
Формализованные правила, селективный режим	0,400	0,322	0,333	0,632	0,000	0,389
TF-IDF + Logistic Regression	0,694	0,682	0,769	0,476	0,800	1,000
TF-IDF + Linear SVM	0,667	0,646	0,769	0,400	0,769	1,000

Лучший результат по средним метрикам показала модель TF-IDF + Logistic Regression: сбалансированная точность составила 0,694 и Macro-F1 = 0,682. Linear SVM также превысила подход на основе формализованных правил по средним метрикам. Вместе с тем подход на основе формализованных правил формирует объяснимое решение, показывает найденные и отсутствующие доказательства и выделяет спорные случаи для экспертной проверки.

Анализ статусов модели на основе формализованных правил на контрольной выборке показал, что из 36 заявок 14 получили окончательное автоматическое решение, 9 были направлены на экспертную проверку, 7 отнесены к случаям с противоречивыми признаками, 5 – к пограничным случаям, 1 – к случаям с недостаточностью данных. Также выявлен 1 случай расхождения решений между языковыми версиями текста.

Обсуждение

Полученные результаты показывают, что формализация содержательных признаков позволяет представить проверку этапа НИОКР как последовательную и объяснимую процедуру. При этом текущая версия модели на основе формализованных правил уступает классическим базовым сравнительным моделям TF-IDF по численным метрикам. Поэтому вклад предложенного подхода в данной работе следует формулировать не как превосходство по точности, а как формализацию экспертных критериев, прозрачность решения и возможность направлять неоднозначные заявки на проверку.

Селективный режим в текущем эксперименте не улучшил качество на покрытой части выборки. Его значение состоит в другом: модель не обязана принимать решение по каждой заявке и может фиксировать пограничные случаи, случаи с противоречивыми признаками, необходимость экспертной проверки и недостаточность данных. Для экспертной системы это важное свойство, поскольку спорные или неполные заявки не должны автоматически получать окончательный класс.

Главное ограничение исследования связано с данными. Класс ОКР представлен только 36 заявками, а пилотная выборка искусственно сбалансирована и не отражает реальные доли классов. Кроме того, research_type является административной меткой, а не независимой экспертной разметкой. Test-часть содержит по 12 заявок на класс, поэтому одна ошибка меняет полноту (recall) класса на 8,33 процентного пункта. В дальнейшем необходимо расширить представительство ОКР, провести ручную экспертную проверку части заявок и отдельно разобрать ошибки между ПИ и ОКР.

Заключение

В работе предложена объяснимая информационно-аналитическая модель классификации этапов НИОКР по тексту конкурсных заявок. Модель использует формальные индикаторы для трёх классов и механизм категориальной оценки уверенности.

Пилотная апробация на 108 обезличенных заявках показала, что подход на основе формализованных правил обеспечивает интерпретируемость и выявляет спорные случаи, но по метрикам качества уступает базовым сравнительным моделям TF-IDF + Logistic Regression и TF-IDF + Linear SVM. Лучшая базовая сравнительная модель достигла Macro-F1 = 0,682 против 0,548 у принудительной классификации на основе формализованных правил.

Практическая значимость подхода состоит в предварительном выявлении возможных несоответствий между заявленным типом исследования и содержанием проекта, а также в прозрачном объяснении оснований классификации. В текущем виде результаты следует рассматривать как пилотную апробацию процедуры поддержки эксперта, а не как готовую промышленную модель.

Информация о финансировании

Данное исследование финансировалось Комитетом науки МНВО РК (программно-целевое финансирование на 2025-2027 годы ИРН BR28713385 «Разработка интеллектуальной системы экспертизы и оценки научных, научно-технических проектов»).

Литература

1. Tennant, J. P. The limitations to our understanding of peer review / J. P. Tennant, T. Ross-Hellauer // Research Integrity and Peer Review. – 2020. – Vol. 5, no. 1. – Art. 6. – DOI: https://doi.org/10.1186/s41073-020-00092-1.
2. Bias in peer review / C. J. Lee, C. R. Sugimoto, G. Zhang, B. Cronin // Journal of the American Society for Information Science and Technology. – 2013. – Vol. 64, no. 1. – P. 2–17. – DOI: https://doi.org/10.1002/asi.22784.
3. Bias in research grant evaluation has dire consequences for small universities / D. Murray, D. Morris, C. Lavoie [et al.] // PLoS ONE. – 2016. – Vol. 11, no. 6. – Art. e0155876. – DOI: https://doi.org/10.1371/journal.pone.0155876.
4. Low agreement among reviewers evaluating the same NIH grant applications / E. L. Pier, M. Brauer, A. Filut [et al.] // Proceedings of the National Academy of Sciences. – 2018. – Vol. 115, no. 12. – P. 2952–2957. – DOI: https://doi.org/10.1073/pnas.1714379115.
5. Об утверждении Правил организации и проведения государственной научно-технической экспертизы: приказ Министра науки и высшего образования Республики Казахстан от 7 ноября 2024 года № 517 : зарегистрирован в Министерстве юстиции

Республики Казахстан 8 ноября 2024 года № 35344 // Информационно-правовая система «Әділет». – URL: https://adilet.zan.kz/rus/docs/V2400035344 (дата обращения: 15.03.2026).
6. A conceptual framework for AI- and blockchain-enabled research project evaluation systems / S. Amanzholova, G. Mutanov, O. Ussatova [et al.] // Information. – 2026. – Vol. 17, no. 2. – Art. 151. – P. 1–19. – DOI: https://doi.org/10.3390/info17020151.
7. International experience and digital platforms for the examination of scientific and technical projects: a critical review and recommendations for Kazakhstan / G. Mutanov, S. Amanzholova, I. Shayea, L. Aldasheva // Труды университета. – 2026. – № 1 (102). – P. 471–478. – DOI: https://doi.org/10.52209/1609-1825_2026_1_471.
8. Президент провел в КазНУ имени аль-Фараби встречу с научной общественностью: выступление Президента Республики Казахстан от 19 марта 2026 года // Официальный сайт Президента Республики Казахстан. – URL: https://www.akorda.kz/ru/prezident-provel-v-kaznu-imeni-al-farabi-vstrechu-s-nauchnoy-obshchestvennostyu-1923227 (дата обращения: 10.09.2026).
9. Глава государства выступил на заседании Национального совета по науке и технологиям: выступление Президента Республики Казахстан от 26 сентября 2025 года // Официальный сайт Президента Республики Казахстан. – URL: https://www.akorda.kz/ru/glava-gosudarstva-vystupil-na-zasedanii-nacionalnogo-soveta-po-nauke-i-tehnologiyam-2684537 (дата обращения: 10.09.2026).
10. О науке и технологической политике: Закон Республики Казахстан от 1 июля 2024 года № 103-VIII ЗРК // Информационно-правовая система «Әділет». – URL: https://adilet.zan.kz/rus/docs/Z2400000103 (дата обращения: 15.08.2026).
11. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development / OECD. – Paris: OECD Publishing, 2015. – 400 p. – DOI: https://doi.org/10.1787/9789264239012-en.
12. Godin, B. The linear model of innovation: the historical construction of an analytical framework / B. Godin // Science, Technology, & Human Values. – 2006. – Vol. 31, no. 6. – P. 639–667. – DOI: https://doi.org/10.1177/0162243906291865.
13. Петров, А. Н. Оценка уровня технологической готовности конкурсных заявок с использованием методологии TPRL / А. Н. Петров, А. В. Комаров // Экономика науки. – 2020. – Т. 6, № 1–2. – С. 88–99.
14. Об утверждении Методики определения уровней готовности технологий и технологической готовности организаций: приказ Министра науки и высшего образования Республики Казахстан от 10 января 2025 года № 8: зарегистрирован в Министерстве юстиции Республики Казахстан 13 января 2025 года № 35634 // Информационно-правовая система «Әділет». – URL: https://adilet.zan.kz/rus/docs/V2500035634 (дата обращения: 05.07.2026).
15. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review / R. Ofori-Boateng, M. Aceves-Martins, N. Wiratunga, C. F. Moreno-García // Artificial Intelligence Review. – 2024. – Vol. 57. – Art. 200. – DOI: https://doi.org/10.1007/s10462-024-10844-w.
16. Recall, risk, and governance in automated proposal screening for research funding: evidence from a national funding programme / C. G. Nagarajappa, M. Koley, A. Kumar [et al.] // arXiv. – 2026. – URL: https://arxiv.org/abs/2602.07869 (дата обращения: 18.05.2026).
17. A sentence-level hierarchical BERT model for document classification with limited labelled data / J. Lu, M. Henchion, I. Bacher, B. Mac Namee // Discovery Science. – Cham: Springer, 2021. – P. 231–241. – DOI: https://doi.org/10.1007/978-3-030-88942-5_18.
18. Managing AI for research excellence: embedding ethics, equity and inclusion: report / Global Research Council Equality, Diversity and Inclusion Working Group. – 2025. – URL: https://globalresearchcouncil.org/fileadmin/documents/GRC_Publications/

GRC_2025_AM_Riyadh_Side_Event_2_Report_-_EDI_WG_Managing_AI_for_Research_Excellence_Embedding_Ethics_Equity_and_Inclusion.pdf (дата обращения: 11.05.2026).

РАЗРАБОТКА ОБЛАЧНОЙ ПЛАТФОРМЫ ДЛЯ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ВЕБ-САЙТОВ И ПРЕЗЕНТАЦИЙ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Нерезов Виктор Константинович
e-mail: 25260222@turan-edu.kz
Университет «Туран», г. Алматы, Казахстан

Аннотация

Создание веб-сайтов и презентаций традиционно требует навыков верстки и дизайна, что ограничивает доступность таких инструментов. В работе представлена облачная SaaS-платформа AIPageStudio, автоматически генерирующая веб-сайты и презентации по текстовому описанию пользователя с применением больших языковых моделей. Система интегрирована с Claude API: запрос преобразуется в структурированный JSON, на основе которого формируется HTML/CSS-разметка сайта или слайды презентации. Архитектура построена на восьми независимых Go-микросервисах с единой точкой входа – API Gateway с JWT-аутентификацией. Пользователь редактирует контент в визуальном drag-and-drop редакторе, публикует сайт по постоянной ссылке и экспортирует результат в ZIP или PPTX. Инфраструктура контейнеризована для развёртывания на Docker и Kubernetes.

Введение

Цифровая трансформация бизнеса и образования формирует устойчивый спрос на инструменты быстрого создания веб-присутствия и визуального представления информации. Однако традиционные инструменты – визуальные редакторы, конструкторы сайтов и редакторы презентаций – требуют от пользователя навыков верстки и дизайна, что существенно ограничивает их доступность широкому кругу пользователей.

Появление больших языковых моделей (LLM) [1, 2] и развитие техник prompt engineering [4–6] показали, что языковые модели способны стабильно генерировать структурированные данные – JSON, HTML, XML, – что открывает возможность их прямой интеграции в производственные системы.

Существующие коммерческие решения – Wix ADI, Webflow, Gamma, Canva – лишь частично решают эту задачу: одни ориентированы только на сайты, другие – только на презентации, и все ограничивают пользователя фиксированным набором шаблонов, не позволяя генерировать контент с нуля по произвольному описанию. Разрыв между «текстовым ответом LLM» и «готовым к публикации веб-продуктом» остаётся незакрытым.

Целью данной работы является разработка облачной SaaS-платформы AIPageStudio, которая устраняет указанный разрыв: пользователь формулирует текстовое описание – система автоматически генерирует готовый веб-сайт или презентацию, доступные для редактирования и экспорта. Платформа построена на микросервисной архитектуре [10, 11], обеспечивающей независимое масштабирование компонентов, и интегрирована с Claude API компании Anthropic в качестве ядра генерации контента.

Обзор существующих подходов и научный контекст

Теоретической основой для разработки платформы послужили исследования в трёх взаимосвязанных областях: архитектура и возможности LLM, техники взаимодействия с языковыми моделями, а также проектирование масштабируемых облачных систем.

Фундаментом современных LLM является архитектура Transformer [1] с механизмом самовнимания (self-attention), позволившим моделям эффективно обрабатывать длинные последовательности и сложные семантические зависимости. Дальнейшее масштабирование архитектуры привело к появлению GPT-3 [2] – модели,

продемонстрировавшей способность решать разнообразные задачи по текстовому описанию без дополнительного обучения (few-shot learning). Обзор Zhao et al. [3] подтверждает, что современные модели демонстрируют устойчивую способность к генерации структурированных данных – ключевое требование для интеграции в производственные системы.

Отдельную область исследований образует prompt engineering – дисциплина формирования инструкций для управления поведением языковых моделей. Schulhoff et al. [4] систематизировали более 50 техник промптинга, среди которых для нашей задачи особенно значимы role prompting и output format constraints, а White et al. [5] предложили каталог повторно используемых паттернов промптов для проектирования AI-сервисов. Xing et al. [6] выделили конвейерный подход «промпт – структурированный ответ – downstream-сервис» как основной архитектурный паттерн AI-native систем. Lewis et al. [7] исследовали технику RAG, а Yao et al. [8] и Wu et al. [9] развили концепцию автономных агентов, использующих внешние инструменты и API.

В части архитектуры облачных систем ориентирами послужили работы Richardson [10] и Newman [11], описывающие паттерны API Gateway и Database per Service, а также Kleppmann [12] и Burns et al. [13] – по надёжности систем с высокой нагрузкой и оркестрации контейнеров Kubernetes. Эти подходы определили архитектурные решения AIPageStudio.

Для систематизации анализа существующих коммерческих решений в таблице 1 приведено их сравнение по ключевым функциональным критериям применительно к задаче автоматической генерации веб-контента.

Платформа	Генерация сайтов	Генерация презентаций	Генерация по произвольному промпту	Визуальный редактор	Экспорт
Wix ADI	+	–	– (шаблоны)	+	–
Webflow	–	–	–	+	HTML/CSS
Gamma	–	+	+	частично	PDF, PPTX
Canva	частично	+	частично	+	PDF, PPTX
AIPageStudio	+	+	+	+	ZIP, PPTX

Таблица 1 – Сравнительный анализ существующих платформ для создания веб-контента

Как следует из таблицы 1, ни одна из платформ не обеспечивает одновременно генерацию обоих типов контента по произвольному описанию с визуальным редактированием и гибким экспортом. При этом работы, описывающие совместную интеграцию LLM-генерации, prompt engineering и микросервисной архитектуры в единую платформу, в открытой литературе практически отсутствуют – на заполнение этого пробела направлена настоящая работа.

Объект, субъект и методы исследования

Объектом исследования является процесс автоматизированного создания веб-контента и презентаций с применением больших языковых моделей. Предметом – архитектурные решения и методы интеграции LLM в облачную SaaS-платформу, обеспечивающие преобразование текстового описания пользователя в готовый, редактируемый веб-продукт. Работа носит прикладной инженерно-исследовательский характер.

Ключевым методическим решением стало разделение системы на независимые микросервисы [10, 11] для изоляции ответственности и независимого масштабирования компонентов. Платформа AIPageStudio реализована в виде восьми сервисов: api-gateway, auth-service, project-service, ai-service, website-service, presentation-service, export-service и storage-service. Каждый сервис разработан на Go по принципам Clean Architecture с

чётким разделением HTTP-обработчика, бизнес-логики и слоя данных. Единой точкой входа служит api-gateway, централизующий JWT-аутентификацию и маршрутизацию, тогда как downstream-сервисы доверяют идентификатору пользователя, переданному через внутренний заголовок.

Центральным методом работы является prompt engineering [4–6]: для задачи генерации веб-сайта разработан системный промпт, задающий модели роль архитектора контента, формат ответа в виде JSON-схемы и примеры допустимых значений (few-shot). Это обеспечивает стабильный возврат структурированного объекта с полями type, title, color_scheme и массивом sections, пригодного для прямой передачи в website-service. Аналогичный подход применяется для презентаций: ответ модели содержит поля theme и массив slides с заголовками и тезисами.

Выбор модели Claude Sonnet 4.6 обусловлен её высокой точностью в задачах структурированной генерации и надёжным следованием инструкциям промпта. В отличие от подхода RAG [7], AIPageStudio использует прямую генерацию контента без внешних источников, поскольку задача состоит в создании нового материала, а не в поиске существующего. Архитектурные паттерны автономных агентов [8, 9] учтены при проектировании ai-service как изолированного компонента, инкапсулирующего взаимодействие с внешним API.

Методология разработки включала четыре этапа: формирование требований и API-контрактов между сервисами; реализация и тестирование backend-сервисов с миграциями базы данных (golang-migrate) и изолированными PostgreSQL-схемами [12]; разработка фронтенда на Vue 3 с редактором GrapesJS для визуального drag-and-drop редактирования HTML; контейнеризация системы средствами Docker и подготовка к оркестрации через Kubernetes [13], что обеспечивает воспроизводимость окружения и готовность к облачному развёртыванию.

Реализация платформы AIPageStudio

С архитектурной точки зрения AIPageStudio – модульная система из трёх уровней: пользовательский интерфейс, бизнес-логика в виде набора микросервисов и хранение данных. Такое разделение позволяет независимо развивать и масштабировать каждый компонент, не затрагивая остальные. На рисунке 1 представлена общая архитектура платформы.



Рисунок 1 – Архитектура платформы AIPageStudio

Ядром системы является ai-service, инкапсулирующий всё взаимодействие с Claude API. Сервис формирует итоговый промпт из системной инструкции и описания пользователя, после чего передаёт его модели. Системный промпт задаёт роль, формат

ответа и ограничения: модель обязана вернуть валидный JSON заданной схемы – для сайта это поля type, title, color_scheme и массив sections, для презентации – theme и массив slides с заголовками и тезисами. После валидации JSON-структуры ai-service возвращает её фронтенду, который вызывает website-service или presentation-service. Такое разделение ответственности позволяет ai-service оставаться изолированным от деталей хранения и рендеринга.

Website-service преобразует полученный JSON в HTML/CSS-разметку с применением Go-шаблонов: каждому типу секции (hero, features, about, contact) соответствует свой шаблон. Готовый HTML сохраняется в базе данных и сразу доступен для визуального редактирования в браузере на основе GrapesJS – open-source библиотеки drag-and-drop редактирования HTML, позволяющей менять содержимое, блоки и стили без знания HTML. Публикация сайта выполняется одним действием и делает его доступным по постоянной ссылке вида /website/p/{project_id}. Экспорт в ZIP реализован в export-service средствами стандартной библиотеки Go archive/zip, без внешних зависимостей.

Аналогичная логика применяется к презентациям: presentation-service хранит слайды в формате JSON и предоставляет API для редактирования заголовков, тезисов и порядка слайдов. Экспорт в PPTX выполняется export-service на основе формата OOXML – ZIP-архив с XML-файлами по спецификации Open Packaging Conventions, собираемый исключительно средствами text/template и archive/zip стандартной библиотеки Go, без зависимости от внешних C-библиотек.

Безопасность реализована через двухуровневую схему токенов: auth-service выдаёт JWT-токен доступа с коротким сроком жизни и refresh-токен, хранящийся в базе данных в виде SHA-256 хэша. Api-gateway – единственный компонент, знающий JWT-секрет: он проверяет подпись и срок токена при каждом запросе и передаёт идентификатор пользователя downstream-сервисам через внутренний заголовок, что снижает поверхность атаки. Файловое хранилище реализовано через storage-service, взаимодействующий с MinIO по S3-совместимому API, что позволяет заменить его облачным бакетом без изменения кода.

Фронтенд разработан как SPA на Vue 3 с Vite, TypeScript и Pinia. Интерфейс поддерживает три языка (казахский, русский, английский) и две цветовые темы, взаимодействуя с backend через api-gateway по REST с JWT-авторизацией. Пользовательский сценарий сведён к минимуму шагов: регистрация – создание проекта – ввод описания – генерация – редактирование – публикация или экспорт. Именно эта целостность сценария – от текстового запроса до готового продукта – отличает AIPageStudio от инструментов, останавливающихся на уровне генерации текста.

Результаты разработки и апробации платформы

В ходе работы была полностью реализована облачная платформа AIPageStudio: восемь backend-микросервисов на Go с единой внутренней структурой, фронтенд-приложение и инфраструктурный слой. Взаимодействие компонентов проверено в составе Docker Compose окружения, включающего все сервисы, PostgreSQL и MinIO. На рисунке 2 представлен интерфейс платформы с результатом генерации веб-сайта.

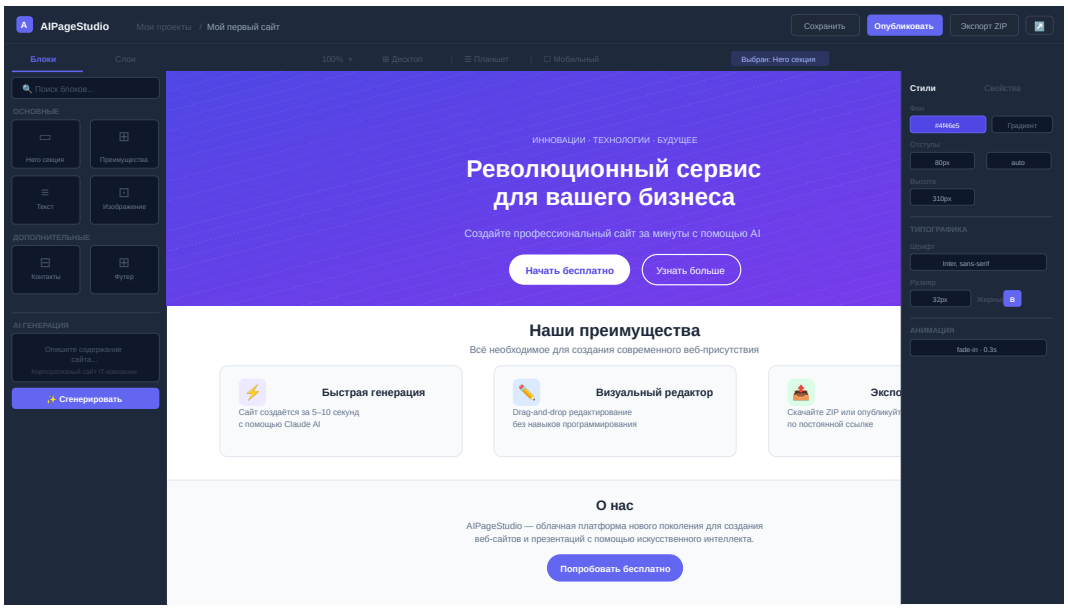


Рисунок 2 – Интерфейс платформы AIPageStudio: результат генерации и визуальный редактор

Ключевым результатом является работоспособный конвейер AI-генерации: система стабильно преобразует текстовое описание в структурированный JSON и далее в готовый HTML-сайт или набор слайдов за 5–10 секунд (в основном время ответа Claude API). Валидация JSON-схемы предотвращает передачу некорректных данных в downstream-сервисы и обеспечивает предсказуемость поведения даже при нестандартных запросах. Модель корректно формирует заголовки, секции и цветовую схему в соответствии с тематикой.

Функционал редактирования, публикации и экспорта реализован в полном объёме. Сгенерированный сайт сразу открывается в GrapesJS-редакторе, позволяя перемещать блоки, редактировать текст и стили. Публикация выполняется одним нажатием и делает сайт доступным по постоянной ссылке. Экспорт в ZIP и PPTX корректно работает для всех проверенных конфигураций; скачиваемые файлы открываются в Microsoft PowerPoint и стандартных архиваторах без ошибок.

Вместе с тем апробация выявила ограничения: качество генерации существенно зависит от детализации промпта – короткие и абстрактные описания дают менее релевантные результаты. Часть пользовательских сценариев (в особенности управление проектами и загрузка изображений) реализована не в полном объёме, а механизм повторной генерации отдельных секций без перегенерации всей страницы пока отсутствует.

Заключение

Проведённая работа показала, что задача автоматической генерации веб-сайтов и презентаций на основе текстового описания может быть успешно решена в формате облачной SaaS-платформы, объединяющей LLM, микросервисную архитектуру и визуальный редактор. Реализованная платформа AIPageStudio – восемь backend-сервисов на Go, фронтенд на Vue 3 и контейнеризованная инфраструктура на Docker и Kubernetes – стабильно преобразует пользовательский промпт в готовый HTML-сайт или PPTX-презентацию за 5–10 секунд; все ключевые компоненты (генерация, хранение, рендеринг, редактирование, публикация, экспорт) реализованы и верифицированы в едином рабочем окружении.

Научная ценность работы состоит в демонстрации инженерного подхода к интеграции LLM в производственную систему – от проектирования JSON-схем и системных промптов до изоляции AI-сервиса в микросервисной архитектуре. Прикладная ценность – в возможности использовать AIPageStudio как инструмент

быстрого прототипирования веб-контента без навыков верстки. Дальнейшее развитие платформы включает расширение библиотеки шаблонов, поддержку частичной перегенерации секций, историю версий и мобильный интерфейс.

ЛИТЕРАТУРА

1. *Vaswani A., Shazeer N., Parmar N. et al.* Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – DOI: 10.48550/arXiv.1706.03762 (дата обращения 24.05.2026).

2. *Brown T., Mann B., Ryder N. et al.* Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 1877–1901. – DOI: 10.48550/arXiv.2005.14165 (дата обращения 24.05.2026).

3. *Zhao W., Zhou K., Li J. et al.* A Survey of Large Language Models // arXiv. – 2023. – DOI: 10.48550/arXiv.2303.18223 (дата обращения 24.05.2026).

4. *Schulhoff S., Ilie M., Zheng S. et al.* The Prompt Report: A Systematic Survey of Prompting Techniques // arXiv. – 2024. – DOI: 10.48550/arXiv.2406.06608 (дата обращения 24.05.2026).

5. *White J., Fu Q., Hays S. et al.* A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT // arXiv. – 2023. – DOI: 10.48550/arXiv.2302.11382 (дата обращения 24.05.2026).

6. *Xing Z., Huang Q., Cheng Y., Zhu L., Lu Q., Xu X.* Prompt Sapper: LLM-Empowered Software Engineering Infrastructure for AI-Native Services // arXiv. – 2023. – DOI: 10.48550/arXiv.2306.02230 (дата обращения 24.05.2026).

7. *Lewis P., Perez E., Piktus A. et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459–9474. – DOI: 10.48550/arXiv.2005.11401 (дата обращения 24.05.2026).

8. *Yao S., Zhao J., Yu D. et al.* ReAct: Synergizing Reasoning and Acting in Language Models // arXiv. – 2022. – DOI: 10.48550/arXiv.2210.03629 (дата обращения 24.05.2026).

9. *Wu Q., Bansal G., Zhang J. et al.* AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation // arXiv. – 2023. – DOI: 10.48550/arXiv.2308.08155 (дата обращения 24.05.2026).

10. *Richardson C.* Microservices Patterns: With Examples in Java. – Shelter Island: Manning Publications, 2018. – 520 p.

11. *Newman S.* Building Microservices: Designing Fine-Grained Systems. – 2nd ed. – Sebastopol: O'Reilly Media, 2021. – 616 p.

12. *Kleppmann M.* Designing Data-Intensive Applications. – Sebastopol: O'Reilly Media, 2017. – 590 p.

13. *Burns B., Beda J., Hightower K., Evenson L.* Kubernetes: Up and Running. – 3rd ed. – Sebastopol: O'Reilly Media, 2022. – 326 p.

УДК 004.75

ОБЛАЧНЫЕ ТЕХНОЛОГИИ КАК ОСНОВА СОВРЕМЕННОЙ IT-ИНФРАСТРУКТУРЫ

Раимбаев Мухаммадамин Самандаругли
Студент
Таразский Региональный Университет имени М. Х. Дулати,
Республика Казахстан, г. Тараз
E-mail: mraimbaev72@gmail.com

Введение

Развитие информационных технологий сопровождается постоянным ростом объёмов обрабатываемых данных, требований к скорости работы информационных систем и необходимости обеспечивать доступность цифровых сервисов для большого количества пользователей. Традиционная модель, при которой организация самостоятельно приобретает серверы, системы хранения данных и сетевое оборудование, постепенно дополняется и во многих случаях заменяется использованием облачной инфраструктуры. Изменение связано не только с появлением новых способов размещения вычислительных ресурсов, но и с переходом к другой модели организации IT-инфраструктуры, в которой вычислительные мощности, хранилища, базы данных, сетевые сервисы и программное обеспечение предоставляются как управляемые услуги.

В соответствии с определением Национального института стандартов и технологий США (NIST), облачные вычисления представляют собой модель сетевого доступа по требованию к общему пулу настраиваемых вычислительных ресурсов, которые могут оперативно предоставляться и освобождаться с минимальными усилиями по управлению. В модели NIST выделяются пять основных характеристик облака, три модели предоставления услуг и четыре модели развертывания [1].

Актуальность исследования обусловлена тем, что облачные технологии сегодня используются не только для хранения файлов или размещения веб-сайтов. На их основе создаются корпоративные информационные системы, платформы электронной коммерции, мобильные приложения, аналитические системы, базы данных, системы искусственного интеллекта и распределённые программные комплексы. Облачная инфраструктура становится связующим уровнем между программным обеспечением и физическими вычислительными ресурсами.

Цель статьи заключается в рассмотрении роли облачных технологий как основы современной IT-инфраструктуры, анализе основных моделей облачных услуг и развертывания, преимуществ и ограничений облачной среды, а также определении её значения для разработки программного обеспечения и дальнейшего развития цифровых систем.

Понятие и сущность облачных технологий

Понятие облачных технологий включает совокупность программных и аппаратных средств, обеспечивающих удалённое предоставление вычислительных ресурсов через сеть. При этом пользователь взаимодействует не с конкретным физическим сервером, а с логическим ресурсом, характеристики которого могут изменяться в зависимости от нагрузки и требований информационной системы.

Сущность облачного подхода заключается в отделении используемого IT-ресурса от конкретного физического оборудования. Например, организации не обязательно приобретать отдельный сервер для размещения веб-приложения. Вместо этого она может арендовать виртуальную вычислительную машину, настроить необходимые параметры процессора, оперативной памяти и дискового пространства, а затем подключить базу данных, систему резервного копирования и средства мониторинга. При увеличении нагрузки ресурсы могут быть увеличены, а после снижения нагрузки — уменьшены.

NIST выделяет пять характеристик, которые позволяют отличать полноценную облачную модель от обычного удалённого размещения оборудования: самообслуживание по требованию, широкий сетевой доступ, объединение ресурсов, быстрая эластичность и измеряемое потребление [1]. Эти характеристики важны именно в совокупности. Например, наличие удалённого сервера само по себе ещё не означает использование полноценной облачной модели.

Облачная инфраструктура строится на нескольких технологических уровнях. На физическом уровне находятся серверы, системы хранения и сетевое оборудование дата-центров. Далее используются технологии виртуализации и программно определяемой инфраструктуры. Следующий уровень образуют вычислительные, сетевые и дисковые сервисы, базы данных, системы мониторинга и средства управления. На верхнем уровне располагаются прикладные программы, которыми непосредственно пользуются сотрудники или клиенты организации.

Таким образом, облако представляет собой не отдельную технологию, а комплексную архитектурную модель. В ней физические ресурсы превращаются в управляемые и масштабируемые сервисы. Именно это свойство позволяет рассматривать облачные технологии не просто как альтернативу локальным серверам, а как один из фундаментальных элементов современной IT-инфраструктуры.

Роль облачных технологий в современной IT-инфраструктуре

Современная IT-инфраструктура должна одновременно обеспечивать вычисления, хранение информации, сетевое взаимодействие, безопасность, резервное копирование, мониторинг и поддержку программных систем. При традиционном подходе все перечисленные компоненты во многом зависят от оборудования конкретной организации. Облачная модель переносит значительную часть инфраструктурных функций в специализированную среду поставщика услуг.

Одной из главных особенностей облачной инфраструктуры является возможность разделения физического оборудования между несколькими клиентами при сохранении логической изоляции их ресурсов. Такой подход позволяет повысить эффективность использования вычислительных мощностей. Виртуальные машины, контейнеры, системы хранения и сетевые ресурсы могут динамически распределяться между рабочими нагрузками.

Важную роль играет масштабируемость. Если веб-сервис рассчитан на несколько тысяч пользователей, а во время определённого события число запросов резко возрастает, традиционная инфраструктура должна заранее иметь достаточный запас серверных мощностей. В облаке часть ресурсов может быть увеличена динамически. После снижения нагрузки дополнительные мощности освобождаются. Именно эластичность является одним из принципиальных свойств облачной модели [1; 7].

Экономическая составляющая также влияет на архитектурные решения. Исследование M. Armbrust и соавторов показывает, что облачная модель позволяет отказаться от необходимости заранее приобретать избыточные вычислительные ресурсы и использовать инфраструктуру в соответствии с текущей потребностью [7]. При этом экономическая эффективность не является автоматической: расходы зависят от архитектуры, объёма хранения, передачи данных, используемых сервисов и продолжительности их работы. Поэтому переход в облако должен сопровождаться анализом совокупной стоимости владения и контролем потребления ресурсов.

Облачная инфраструктура также упрощает создание резервных копий и организацию аварийного восстановления. Данные и сервисы могут размещаться в нескольких зонах доступности или регионах, а восстановление может выполняться автоматически или по заранее определённым сценариям. Однако высокая доступность не возникает только благодаря факту использования облака. Архитектура приложения должна изначально учитывать отказ отдельных компонентов, резервирование и процедуры восстановления [4; 9].

Следовательно, значение облачных технологий заключается не только в переносе серверов из локального помещения в дата-центр поставщика. Главное изменение состоит в переходе от управления отдельными физическими устройствами к управлению набором программно определяемых ресурсов и сервисов.

Модели облачных услуг

В соответствии с классификацией NIST выделяются три основные модели предоставления облачных услуг: Infrastructure as a Service, Platform as a Service и Software as a Service [1]. Эти модели отличаются уровнем контроля пользователя над инфраструктурой.

Infrastructure as a Service (IaaS) предоставляет базовые вычислительные ресурсы. Пользователь получает виртуальные машины, хранилища, сетевые компоненты и другие инфраструктурные средства, а операционные системы и прикладное программное обеспечение настраивает самостоятельно. IaaS подходит организациям, которым требуется значительный контроль над программной средой. Например, компания может развернуть виртуальный сервер Linux, установить веб-сервер Nginx, приложение на Python или Java, систему управления базами данных и необходимые средства мониторинга.

Platform as a Service (PaaS) располагается на более высоком уровне абстракции. Пользователю не требуется самостоятельно управлять значительной частью операционной системы и инфраструктуры. Основное внимание разработчик уделяет программному коду и данным. Такая модель удобна при создании веб-приложений и API, поскольку разработчик может передать платформе часть задач, связанных с масштабированием, развертыванием и обслуживанием среды выполнения [5].

Software as a Service (SaaS) предполагает предоставление готового приложения через сеть. Пользователь не занимается установкой и обслуживанием серверной инфраструктуры. К этому классу относятся различные корпоративные системы, средства совместной работы, системы управления взаимоотношениями с клиентами, электронная почта и другие приложения, предоставляемые как сервис.[2,5]

Преимущества и ограничения облачных технологий

Одним из наиболее существенных преимуществ облачных технологий является скорость получения IT-ресурсов. При локальном размещении нового сервера необходимо приобрести оборудование, установить его, подключить к сети, настроить операционную систему и программное обеспечение. В облачной среде аналогичная вычислительная среда может быть создана программными средствами значительно быстрее.

Другим преимуществом является масштабируемость. Для интернет-магазина, например, количество пользователей может резко изменяться в зависимости от рекламных кампаний, праздников или сезонного спроса. Облачная инфраструктура позволяет использовать автоматическое масштабирование вычислительных ресурсов. Аналогичный принцип применяется к API, мобильным приложениям, системам обработки изображений и аналитическим платформам.

Облако также предоставляет широкий выбор управляемых сервисов. Организация может использовать управляемые базы данных, очереди сообщений, объектные хранилища, системы мониторинга, средства машинного обучения и другие компоненты, не создавая каждый из них самостоятельно. Это сокращает объём инфраструктурной работы и позволяет разработчикам сосредоточиться на бизнес-логике.

Вместе с тем облачные технологии имеют ограничения. Одной из наиболее значимых проблем является информационная безопасность. Перемещение данных в облако означает изменение границ доверия: часть инфраструктуры и процессов управления находится у внешнего поставщика. NIST указывает на необходимость

учитывать риски, связанные с конфиденциальностью, контролем над данными, доступом и зависимостью от поставщика [3].

Особое значение имеет защита персональных данных. Организация должна понимать, где физически размещаются данные, какие механизмы шифрования используются, кто имеет административный доступ и каким образом выполняется удаление или перенос информации. Для государственных, финансовых, медицинских и других чувствительных систем дополнительно возникают требования законодательства и отраслевого регулирования.

Другой риск связан с vendor lock-in — зависимостью от конкретного поставщика. Если приложение активно использует уникальные управляемые сервисы одной платформы, его перенос на другую инфраструктуру может потребовать существенной переработки. NIST отдельно рассматривает вопросы переносимости и интероперабельности как важные факторы облачной архитектуры [4].

Существует и зависимость от сетевого соединения. При использовании SaaS или удалённой базы данных отсутствие доступа к сети непосредственно влияет на возможность работы пользователей. Поэтому для критически важных систем необходимо предусматривать резервные каналы связи, локальное кэширование, отказоустойчивость и процедуры восстановления.

Таким образом, преимущества облака не следует рассматривать изолированно от рисков. Экономия, масштабируемость и скорость разработки должны сопоставляться с требованиями к безопасности, стоимости передачи данных, совместимости и контролю над инфраструктурой.

Облачные технологии в разработке программного обеспечения

Облачная инфраструктура существенно изменила процесс разработки программного обеспечения. Современное веб-приложение может состоять из frontend-компонента, backend API, базы данных, файлового хранилища, системы аутентификации, очередей сообщений и средств мониторинга. Все эти элементы могут быть размещены в облаке как отдельные управляемые компоненты.

Например, мобильное приложение интернет-магазина может обращаться к облачному API. Серверная часть принимает запросы пользователей, обращается к облачной базе данных, получает информацию о товарах и возвращает результат мобильному клиенту. Изображения товаров могут храниться в объектном хранилище, а система аналитики — обрабатывать события пользователей. При увеличении количества клиентов вычислительные ресурсы API могут масштабироваться отдельно от базы данных и хранилища.

Для разработки корпоративной системы можно использовать IaaS, PaaS или комбинацию различных сервисов. При этом облако позволяет создавать отдельные среды разработки, тестирования и эксплуатации. Разработчик может экспериментировать с новой версией приложения, не изменяя рабочую систему.

Облачный подход особенно тесно связан с DevOps. Его основная идея заключается в объединении процессов разработки и эксплуатации, автоматизации доставки программного обеспечения и постоянном контроле состояния системы. Облачная инфраструктура предоставляет удобную основу для реализации CI/CD-процессов, при которых после изменения исходного кода автоматически выполняются сборка, тестирование и развертывание приложения.

Такой подход меняет роль инфраструктуры. Сервер перестаёт восприниматься как уникальный физический объект, который необходимо вручную настраивать. Конфигурация может описываться программным кодом и автоматически применяться к среде. В результате инфраструктура становится частью процесса разработки.

Виртуализация, контейнеризация и автоматизация

Важнейшими технологическими основами облачной инфраструктуры являются виртуализация и контейнеризация. Виртуализация позволяет запускать несколько

изолированных виртуальных машин на одном физическом сервере. Каждая виртуальная машина может иметь собственную операционную систему и набор программ.

Контейнеризация использует другой уровень изоляции. Контейнер включает приложение и необходимые зависимости, что позволяет обеспечить более предсказуемое поведение программного компонента при переносе между средами. Документация Docker характеризует контейнеры как изолированные процессы, содержащие всё необходимое для работы соответствующего компонента приложения [10].

Контейнерный подход особенно важен для микросервисной архитектуры. Большое приложение может быть разделено на отдельные сервисы: авторизация, каталог товаров, обработка заказов, платежи и уведомления. Каждый сервис может разрабатываться и развёртываться независимо.

Для управления большим количеством контейнеров применяется оркестрация. Kubernetes предоставляет платформу для автоматизации развертывания, масштабирования и управления контейнеризированными приложениями [11]. Он позволяет описывать желаемое состояние системы и автоматически поддерживать его, распределять нагрузки между узлами, выполнять обновления и восстанавливать компоненты при отказах.

Практическая значимость облачных технологий для организаций

Практическая ценность облачных технологий определяется возможностью применять их на разных уровнях деятельности организации. Для небольшой компании облако может заменить необходимость создания собственного серверного помещения. Для крупного предприятия оно может использоваться как дополнительный вычислительный ресурс, среда резервного копирования, аналитическая платформа или основа гибридной инфраструктуры.

В разработке веб-приложений облако позволяет быстро создать серверную среду, подключить базу данных и организовать автоматическое масштабирование. В мобильных приложениях облачные сервисы используются для хранения пользовательских данных, синхронизации, авторизации, обработки уведомлений и аналитики. В корпоративных информационных системах облачные платформы могут обеспечивать базы данных, документооборот, системы управления ресурсами и интеграционные API.

Для аналитики данных облако позволяет создавать распределённые хранилища и вычислительные среды без необходимости самостоятельно приобретать вычислительный кластер. Для систем искусственного интеллекта облачная инфраструктура предоставляет доступ к специализированным ускорителям и управляемым платформам машинного обучения.

При этом внедрение облака должно начинаться не с выбора конкретного поставщика, а с анализа требований организации. Необходимо определить типы данных, требования к доступности, допустимые задержки, требования законодательства, модель угроз, предполагаемую нагрузку и бюджет. После этого выбираются модель услуги и модель развертывания.

Важное значение имеет архитектурное управление. Microsoft Cloud Adoption Framework, например, рассматривает облачное внедрение как последовательный процесс, включающий стратегию, планирование, подготовку, миграцию, модернизацию, управление и безопасность [15].

Следовательно, облачная трансформация не должна сводиться к механическому переносу существующих серверов. В ряде случаев перенос старой архитектуры без изменений не даст ожидаемого эффекта. Наиболее значительный результат возникает тогда, когда организация адаптирует архитектуру приложений к возможностям облачной

среды: автоматическому масштабированию, управляемым сервисам, контейнеризации, мониторингу и автоматизированному развертыванию.

Заключение

Облачные технологии стали одним из фундаментальных компонентов современной IT-инфраструктуры, поскольку позволяют представить вычислительные мощности, хранение данных, сетевые ресурсы и программные платформы в форме управляемых сервисов. Их значение определяется не отдельным преимуществом, а совокупностью характеристик: самообслуживанием, эластичностью, масштабируемостью, сетевой доступностью и возможностью автоматизированного управления ресурсами [1].

Анализ моделей IaaS, PaaS и SaaS показывает, что облачные технологии позволяют распределять ответственность между поставщиком и пользователем в зависимости от требуемого уровня контроля. Публичные, частные, гибридные и мультиоблачные архитектуры, в свою очередь, позволяют выбирать способ размещения ресурсов в соответствии с требованиями конкретной организации.

При этом использование облака не устраняет технические риски. Информационная безопасность, защита персональных данных, зависимость от поставщика, сетевые сбои, переносимость данных и совместимость систем остаются важными задачами. Поэтому эффективная облачная инфраструктура должна проектироваться с учётом отказоустойчивости, резервного копирования, контроля доступа, шифрования, мониторинга и заранее определённых процедур восстановления.

Особое значение облачные технологии приобрели в разработке программного обеспечения. Их развитие тесно связано с DevOps, CI/CD, виртуализацией, контейнеризацией, Kubernetes и cloud-native архитектурой. Благодаря этому инфраструктура постепенно превращается из набора отдельных физических устройств в программно управляемую среду, способную изменяться вместе с потребностями приложения.

Перспективы развития связаны с интеграцией облачных вычислений с искусственным интеллектом, Big Data, Internet of Things и edge computing. При этом облачная инфраструктура не обязательно заменяет локальные системы. Более вероятной архитектурной моделью является их взаимодействие, при котором вычисления распределяются между центральным облаком, корпоративной инфраструктурой и периферийными устройствами в зависимости от требований к производительности, задержке, безопасности и стоимости.

Таким образом, облачные технологии можно рассматривать как основу современной IT-инфраструктуры не потому, что они универсально заменяют все традиционные решения, а потому, что они формируют гибкий технологический слой, на котором могут строиться современные программные и информационные системы. Эффективность такой инфраструктуры определяется правильным выбором архитектуры, контролем рисков и соответствием используемых облачных сервисов реальным требованиям организации.

Список использованной литературы

1. Mell P., Grance T. The NIST Definition of Cloud Computing. — Gaithersburg: National Institute of Standards and Technology, 2011. — NIST Special Publication 800-145. — DOI: 10.6028/NIST.SP.800-145.
2. Badger M. L., Grance T., Patt-Corner R., Voas J. Cloud Computing Synopsis and Recommendations. — Gaithersburg: National Institute of Standards and Technology, 2012. — NIST Special Publication 800-146. — DOI: 10.6028/NIST.SP.800-146.
3. Jansen W., Grance T. Guidelines on Security and Privacy in Public Cloud Computing. — Gaithersburg: National Institute of Standards and Technology, 2011. — NIST Special Publication 800-144. — DOI: 10.6028/NIST.SP.800-144.

4. Liu F., Tong J., Mao J., Bohn R., Messina J., Badger M., Leaf D. NIST Cloud Computing Reference Architecture. — Gaithersburg: National Institute of Standards and Technology, 2011. — NIST Special Publication 500-292.
5. Simmon E. Evaluation of Cloud Computing Services Based on NIST SP 800-145. — Gaithersburg: National Institute of Standards and Technology, 2018. — NIST Special Publication 500-322. — DOI: 10.6028/NIST.SP.500-322.
6. Buyya R., Broberg J., Goscinski A. Cloud Computing: Principles and Paradigms. — Hoboken: John Wiley & Sons, 2011. — DOI: 10.1002/9780470940105.
7. Armbrust M., Fox A., Griffith R., Joseph A. D., Katz R., Konwinski A., Lee G., Patterson D., Rabkin A., Stoica I., Zaharia M. A View of Cloud Computing // Communications of the ACM. — 2010. — Vol. 53, No. 4. — P. 50–58. — DOI: 10.1145/1721654.1721672.
8. Amazon Web Services. AWS Well-Architected Framework. — AWS Documentation. — 2026.
9. Google Cloud. Google Cloud Well-Architected Framework. — Google Cloud Documentation. — 2026.
10. Docker. What is a container? — Docker Documentation. — 2026.
11. Kubernetes. Kubernetes Documentation: Overview and Concepts. — Kubernetes Documentation. — 2026.
12. Cloud Native Computing Foundation. Cloud Native Definition v1.0. — CNCF, 2019.
13. NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. — Gaithersburg: National Institute of Standards and Technology.
14. NIST. Fog Computing Conceptual Model: Recommendations of the National Institute of Standards and Technology. — Gaithersburg: National Institute of Standards and Technology, 2018. — NIST Special Publication 500-325. — DOI: 10.6028/NIST.SP.500-325.
15. Microsoft. Cloud Adoption Framework for Azure. — Microsoft Learn. — 2026.

ИССЛЕДОВАНИЕ И РАЗРАБОТКА ОНЛАЙН-ПЛАТФОРМЫ ДЛЯ ПРОВЕДЕНИЯ ТЕСТИРОВАНИЙ И ОПРОСОВ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Сатыбалдин А.Т.
e-mail: 25261840@turan-edu.kz
Университет «Туран», г. Алматы, Казахстан

Аннотация. Представлена онлайн-платформа Quiz для тестирований и опросов с интеграцией больших языковых моделей. Решение объединяет синхронные WebSocket-сессии, асинхронное прохождение, автоматическую генерацию заданий и персональные рекомендации по ошибкам. Гибридная архитектура включает PHP, MySQL 8, Node.js и внешний REST API языковой модели. Платформа развёрнута на eee.kz как SaaS-сервис. Результат работы — реализация единого цикла подготовки заданий, проведения тестирования и формирования индивидуальной обратной связи.

Ключевые слова: онлайн-тестирование, WebSocket, машинное обучение, большие языковые модели, генерация заданий, персонализированная обратная связь, SaaS.

Введение и обзор исследований

Развитие дистанционного обучения усиливает потребность в интерактивном контроле знаний и персонализированной обратной связи. Большие языковые модели (LLM) рассматриваются как средство автоматизации подготовки заданий и сопровождения обучающихся [1]. При этом применение генеративного ИИ предполагает сохранение контроля преподавателя над качеством содержания [2]. Исследования адаптивных образовательных платформ и обзоры методов персонализации показывают возможности машинного обучения для управления учебным контентом и обратной связью [3], [4].

Для автоматической генерации вопросов исследуются конвейеры на основе LLM с управлением уровнем таксономии Блума [5] и получением заданий, близких к составленным экспертами [6]. Обзор методов генерации вопросов отмечает проблемы разнообразия структуры и контроля сложности [8]. В области синхронного тестирования платформа EduQuest демонстрирует применение WebSocket и изоляции состояния участников [7], а работа [9] рассматривает контекстные персонализированные подсказки в интеллектуальных репетиторских системах.

Цель работы — исследование подходов и разработка онлайн-платформы для проведения тестирований и опросов в реальном времени с интеграцией методов машинного обучения. Основной вклад состоит в совместной реализации синхронного тестирования, LLM-генерации заданий и рекомендаций по индивидуальным ошибкам. Для достижения цели выполнены анализ аналогов, проектирование гибридной архитектуры, разработка WebSocket-подсистемы, интеграция LLM и развёртывание SaaS-сервиса.

Методы исследования и архитектура платформы

Объект исследования — процесс онлайн-тестирования и опросов в образовательной и корпоративной среде. Предмет исследования — архитектурные и функциональные решения платформы Quiz. Использовались структурный и функциональный анализ, сравнение аналогов, объектно-ориентированное проектирование, итеративная разработка с прототипированием интерфейсов, интеграционное и нагрузочное тестирование подсистемы реального времени. SaaS-модель обеспечивает доступ через браузер и централизованное обновление функциональности.

Архитектура разделяет бизнес-логику, проведение синхронных сессий и интеллектуальные функции. PHP-бэкенд выполняет аутентификацию, операции

создания, чтения, изменения и удаления тестов (CRUD), обработку статистики и управление подписками. Node.js-сервис отвечает за WebSocket-соединения; языковая модель вызывается из PHP через REST API. Данные хранятся в MySQL 8. Основные компоненты приведены в таблице 1.

Таблица 1

Технологический стек платформы Quiz

Компонент	**Технологии**	**Назначение**
Бизнес-логика	PHP, Composer	Авторизация, тесты, статистика, подписки, администрирование
Хранение данных	MySQL 8, InnoDB, JSON	Тесты, вопросы, ответы, пользователи и платежи
Реальное время	Node.js, ws, mysql2	WebSocket-сессии, пул портов, подключение по PIN
ИИ-компонент	REST API LLM	Генерация заданий и рекомендаций по ошибкам
Интерфейс	HTML, CSS, JavaScript, jQuery, Bootstrap, CKEditor, i18next	Редактор тестов, панели пользователей, экраны прохождения
Сервисные функции	PayBox, reCAPTCHA, CSRF, TCPDF	Платежи, защита форм, экспорт результатов в PDF

Модель данных использует JSON для представления пяти типов заданий. Поле questions таблицы quizzes хранит массив объектов с текстом вопроса, вариантами и правильными ответами, типом, временем, баллами и рекомендацией. При создании синхронной сессии структура теста передаётся в виде JSON-снимка. Результаты сохраняются раздельно: online_sessions — для синхронных сессий, offline_answers — для асинхронных опросов, test_answers — для оцениваемых тестов.

Подсистема реального времени и режимы тестирования

Подсистема реального времени реализована отдельным процессом Node.js. Для каждой активной сессии создаётся экземпляр WebSocket.Server на свободном порту из пула freePorts/busyPorts. Выделение отдельных экземпляров разграничивает состояния параллельных сессий. Управляющий HTTP-API использует порт 3000 и доступен только с localhost.

Протокол разделяет события участников и ведущего. Участники отправляют join, rejoin, leave и answer, ведущий — start и next. Сервер передаёт участникам question, started, quiz_end и force_quiz_end, а ведущему — new_client, client_leave, new_answer и answers. Для восстановления соединения предусмотрено событие rejoin с одноразовым токеном участника. Подключение к синхронному тестированию выполняется по шестизначному PIN-коду; ведущий управляет сессией из веб-интерфейса или мобильного браузера.

Платформа поддерживает синхронный опрос или викторину, асинхронный опрос по ссылке и оцениваемый тест с подсчётом баллов. Реализованы вопросы с одним или несколькими правильными ответами, задания на соответствие и последовательность, а также свободный ответ. В lite-режиме участник видит буквы вариантов ответа, а полные формулировки выводятся на общее табло. Компонент табло работает в браузере, в том числе на Smart TV. Эти режимы позволяют использовать платформу как для аудиторных занятий, так и для самостоятельного прохождения.

Генерация заданий и персонализированная обратная связь

Первая интеллектуальная функция — генерация заданий по параметрам преподавателя. Пользователь задаёт тему, количество и тип вопросов; PHP-бэкенд формирует структурированный промпт и отправляет его языковой модели через REST API. Из ответа регулярными выражениями извлекаются вопрос, варианты и правильный

ответ, после чего задания добавляются в тест. Конвейер «тема → промпт → LLM → парсинг → готовые вопросы» встроен в редактор теста (рисунок 1).

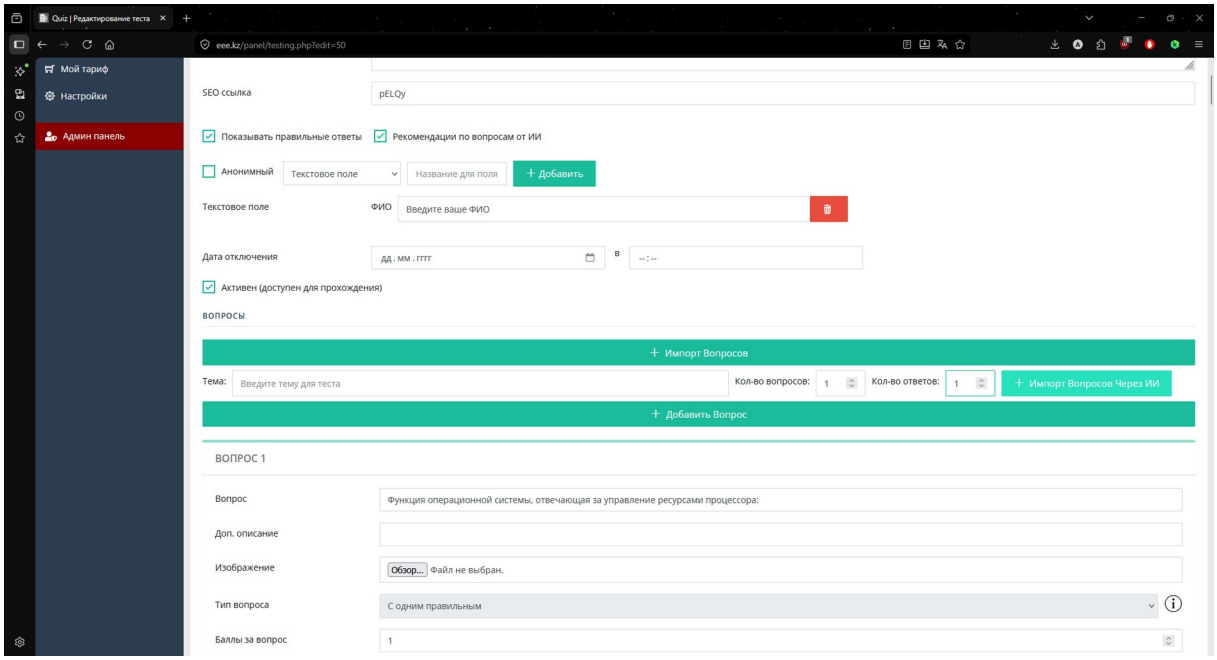


Рисунок 1 — Интерфейс автоматической генерации тестовых заданий

На этапе апробации использованы GPT-4o, GPT-4o-mini и GPT-3.5-turbo через API OpenAI. Выбор модели, ключ API и параметры промпта вынесены в административную панель, поэтому их настройка не требует изменения кода. Собственная специализированная модель относится к дальнейшему развитию, а реализованный вариант использует внешний LLM-сервис. Предусмотрено последующее дообучение на предметных данных с учётом казахстанского образовательного контекста и русско-казахского двуязычия.

Вторая функция — персонализированные рекомендации по неверным ответам. Она активируется параметром recommendation_ai при включённом показе правильных ответов. Языковой модели передаются текст вопроса, варианты и ответ обучающегося; на их основе формируется индивидуальная обратная связь. Если ИИ-функция отключена, выводится статическая рекомендация преподавателя. Таким образом, генерация контента и анализ ошибок включены в один рабочий процесс; подход к контекстным подсказкам соотносится с исследованием [9].

Результаты разработки и обсуждение

Результатом работы стала веб-платформа Quiz, развёрнутая на сервере eee.kz. В ней объединены три режима проведения активностей, пять типов вопросов, WebSocket-управление, генерация заданий и рекомендации. Многоролевая модель включает организацию, частное лицо, преподавателя, социальные опросы и студента; доступ к административной панели задаётся флагом admin. Поддерживаются встраивание через iframe и экспорт результатов в PDF.

Подписки определяют лимиты тестов, вопросов, вариантов ответа и участников. Бесплатный тариф предусматривает 2 теста, 10 вопросов и 100 участников; PRO — 100 тестов, 30 вопросов и неограниченное число участников; «Премиум» — отсутствие лимитов по указанным параметрам. PayBox обрабатывает оплату с серверной проверкой платёжного уведомления.

Функциональное сравнение охватывает Kahoot, Mentimeter и Quizizz. Критериями выбраны режимы проведения, ИИ-инструменты, роли пользователей, встраивание, геймификация, подписки и настраиваемость. Таблица 2 воспроизводит оценки проведённого в работе сравнения; они характеризуют рассмотренную выборку, а не количественную эффективность или текущее состояние всех версий сервисов.

Таблица 2

Функциональное сравнение платформ по материалам исследования

Критерий	**Kahoot**	**Mentimeter**	**Quizizz**	**Quiz**
Реальное время	Да	Да	Да	Да
Асинхронный режим	Нет	Ограниченно	Да	Да
ИИ-генерация заданий	Нет	Нет	Нет	Да
Персональные рекомендации	Нет	Нет	Ограниченно	Да
Многоролевая модель	Ограниченно	Нет	Да	Да
Встраивание через iframe	Нет	Да	Нет	Да
Геймификация	Да	Нет	Да	Да
SaaS и подписки	Да	Да	Да	Да
Открытость и настраиваемость	Нет	Нет	Нет	Да

Сравнение выделяет сочетание генерации заданий и персональных рекомендаций как отличительную возможность Quiz в принятой системе критериев. Его следует рассматривать как качественное сопоставление функций: дата и версии аналогов, методика измерений и численные результаты нагрузочных испытаний в исследовании не приведены. Развёртывание подтверждает наличие работающей реализации, но само по себе не устанавливает масштабируемость или педагогическую эффективность.

Для защиты применены подготовленные SQL-выражения в большинстве моделей, CSRF-токены в сессионных формах и ограничение доступа к управляющему API. В реализации используется MD5-хэш quiz_sign, вычисляемый по идентификатору теста, TOKEN_OFFLINE и вопросам. Эти механизмы описывают текущую реализацию и не означают завершённости защиты. В дальнейшие задачи включены замена MD5 на bcrypt для паролей, параметризация всех SQL-запросов и перенос секретов из config.php в переменные окружения.

Заключение

Разработана и развёрнута SaaS-платформа, объединяющая синхронное WebSocket-тестирование с динамическим выделением сессий, асинхронные режимы, генерацию заданий средствами LLM и обратную связь по ошибкам обучающегося. Основной результат — интеграция подготовки контента, проведения тестирования и персональных рекомендаций в одной системе с многоролевым доступом и подписками.

Дальнейшее развитие предусматривает специализированную языковую модель, адаптивный подбор сложности, аналитику и прогнозирование успеваемости методами классификации и кластеризации, оценку свободных ответов средствами обработки естественного языка, масштабирование через Redis pub/sub и кластеризацию WebSocket-серверов, а также усиление безопасности. Эти направления отделены от уже реализованных функций платформы.

ЛИТЕРАТУРА

1. *Kasneci E., Seßler K., Küchemann S. et al.* ChatGPT for good? On opportunities and challenges of large language models for education // Learning and Individual Differences. — 2023. — Vol. 103. — Art. 102274. — DOI: 10.1016/j.lindif.2023.102274

2. *Giannakos M., Azevedo R., Brusilovsky P. et al.* The promise and challenges of generative AI in education // Behaviour & Information Technology. — 2024. — P. 1–27. — DOI: 10.1080/0144929X.2024.2394880

3. *Sayed W.S., Noeman A.M., Abdellatif A. et al.* AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging e-learning platform // Multimedia Tools and Applications. — 2022. — Vol. 81. — P. 23141–23173. — DOI: 10.1007/s11042-022-13076-8

4. *Al-Mashhadani Z.I., Alrikabi H.T.* Personalized adaptive learning technologies based on machine learning techniques to identify learning styles: A systematic literature review // IEEE Access. — 2023. — DOI: 10.1109/ACCESS.2023.3273308

5. *Scaria N., Chenna S.D., Subramani D.* Automated educational question generation at different Bloom's skill levels using large language models // Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science. — Springer, 2024. — P. 165–179. — DOI: 10.1007/978-3-031-64302-6_12

6. *Wang Z., Valdez J., Basu Mallick D., Baraniuk R.G.* Towards human-like educational question generation with large language models // Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science. — Springer, 2022. — Vol. 13355. — P. 153–166. — DOI: 10.1007/978-3-031-11644-5_13

7. EduQuest: A Hybrid AI-Powered Intelligent Quiz Generation and Real-Time Assessment Platform // International Journal for Research in Applied Science and Engineering Technology. — 2026. — Vol. 14. — DOI: 10.22214/ijraset.2026.77004

8. *Kurdi G., Leo J., Parsia B., Sattler U., Al-Emari S.* A systematic review of automatic question generation for educational purposes // International Journal of Artificial Intelligence in Education. — 2020. — Vol. 30. — P. 121–204. — DOI: 10.1007/s40593-019-00186-y

9. *McNichols H., Liu A., Wang A., Lan A.* Generating in-context, personalized feedback for intelligent tutors with large language models // International Journal of Artificial Intelligence in Education. — 2025. — DOI: 10.1007/s40593-025-00505-6

COURSERA ОНЛАЙН-КУРСТАРЫН ДЕРБЕСТЕНДІРІП ТАҢДАУҒА АРНАЛҒАН ИНТЕЛЛЕКТУАЛДЫ ҰСЫНЫМ ЖҮЙЕСІН ӘЗІРЛЕУ

Д.Н.Исабаева, Д.Н.Акупова
e-mail: Daraja_78@mail.ru; dakupova04@gmail.com
AlmaU, Онлайн білім беру басқармасы, Алматы, Қазақстан

Аңдатпа

Бұл жұмыста университет пәніне сәйкес Coursera онлайн-курстарын автоматты түрде таңдауға арналған интеллектуалды ұсыным жүйесі ұсынылады. Екі Coursera каталогын біріктіру нәтижесінде Course және Short Course санаттары бойынша 20 438 курс қамтылған деректер жиыны қалыптастырылды. Ұсыным алгоритмі multilingual-e5-base моделі арқылы семантикалық іздеу жүргізіп, 80 кандидатты іріктейді, кейін bge-reranker-v2-m3 моделі оларды қайта ранжирлеп, ең сәйкес бес курсты ұсынады. Нәтижелерді іріктеу кезінде курс тілі, субтитрлер мен машиналық аударма сияқты тілдік метадеректер ескеріледі. Төрт ағылшын тіліндегі сұраныс бойынша жүргізілген бастапқы ішкі бағалауда 20 ұсыныстың 19-ы релевантты деп танылып, орташа Precision@5 көрсеткіші 0,95 болды. Ұсынылған прототип университеттік пәндерге сәйкес онлайн-курстарды іздеу мен таңдауды жеңілдетуге бағытталған.

Түйін сөздер: білім берудегі ұсыным жүйесі, Coursera, семантикалық іздеу, көптілді мәтіндік векторлар, қайта ранжирлеу, тілдік қолжетімділік.

Кіріспе

Жаппай ашық онлайн-курстарды (Massive Open Online Courses, MOOC) университет пәніне енгізу үшін курстың танымалдығын ғана емес, оның тақырыптық мазмұнын, оқыту деңгейін және тілдік мүмкіндіктерін салыстыру қажет. Пән атауы каталогтағы курс атауымен дәл сәйкес келмеуі мүмкін: бір тақырып әртүрлі терминдермен сипатталады, ал бірдей термин әр пәндік салада басқа мағынада қолданылады. Сондықтан курс іздеу міндеті кілт сөзді табумен шектелмей, сұраныс контексі мен оқу ресурсының мазмұны арасындағы сәйкестікті бағалауды талап етеді.

Білім берудегі ұсыным жүйелеріне арналған шолуларда мазмұнға негізделген, коллаборативті және гибридті тәсілдер қарастырылады [1]. Коллаборативті тәсіл үшін пайдаланушы–ресурс өзара әрекеттесу деректері қажет. Оқытушы жүйені алғаш қолданғанда мұндай тарих болмауы мүмкін. Осы зерттеуде ақпараттық қажеттілік пән атауы, қосымша сипаттама және таңдалған тіл арқылы беріледі. Мұндағы дербестендіру сұраныс контексіне бейімделуді білдіреді; оқытушының ұзақмерзімді мінез-құлық профилі модельденбейді.

Зерттеудің мақсаты — Coursera каталогынан университет пәніне мазмұндық тұрғыдан сәйкес курстарды іріктейтін екі сатылы жүйені әзірлеу және оның бастапқы нәтижелерін әдістемелік тұрғыдан бағалау. Зерттеу сұрағы: пайдаланушы тарихы жоқ жағдайда көптілді векторлық іздеу, жұптық қайта ранжирлеу және тілдік метадеректерді біріктіру пәнге сәйкес қысқа ұсыным тізімін қалыптастыруға қаншалықты жарамды? Жұмыс дайын нейрондық модельдерді қолданбалы міндетке бейімдейді; жаңа оқыту алгоритмі немесе модель архитектурасы ұсынылмайды.

Әдеби шолу және әдісті таңдау

Urdaneta-Ponte және әріптестері білім берудегі ұсыным жүйелерінің әдістері мен қолданылу бағыттарын жүйелейді [1]. Da Silva және әріптестерінің шолуында бағалаудың көбіне ұсыным дәлдігіне бағытталатыны, ал педагогикалық тиімділікті тексеретін зерттеулердің аз екені көрсетілген [2]. Algarni және Sheldon курс таңдау жүйелерін жеке зерттеу нысаны ретінде қарастырады [3]. Бұл еңбектерден шығатын әдістемелік талап — тақырыптық релеванттылықты оқу нәтижелеріне әсермен теңестірмеу.

Sentence-BERT мәтіндерді бөлек кодтап, векторлық ұқсастық арқылы салыстыру тәсілін негіздейді [4]. Мәтіндерді бөлек кодтайтын мұндай құрылым (bi-encoder) каталог векторларын алдын ала есептеуге мүмкіндік береді. Dense Passage Retrieval жұмысы тығыз векторлық іздеудің сұраққа жауап беру корпусындағы нәтижелерін көрсетеді [5], бірақ ол нәтижелерді білім беру каталогына тікелей көшіруге болмайды. Осы міндет үшін қазақша, орысша және ағылшынша мәтіндермен жұмыс істеуге арналған multilingual E5 моделі таңдалды [6, 7].

Іздеудің екінші кезеңінде мәтіндер жұбын бірге кодтайтын модель (cross-encoder) сұраныс пен кандидат мәтінін бірлесіп өңдейді. BAAI/bge-reranker-v2-m3 құжаттамасы бұл модельдің вектор емес, жұптың релеванттылық балын қайтаратынын нақтылайды [8]. Сондықтан E5 және BGE reranker бірін-бірі алмастыратын екі векторлық кодтау моделі ретінде емес, кандидаттарды жылдам іздеу және оларды қайта реттеу кезеңдері ретінде қолданылады. Осы екі сатылы тәсілдің нақты сапа өсімі бөлек салыстырмалы экспериментті қажет етеді.

Деректер және есептің қойылымы

Жүйеге Coursera каталогының екі Excel файлы енгізілген: негізгі каталог және AI Curriculum for Universities жиынтығы. Негізгі каталогта курс атауы, сипаттамасы, дағдылары, серіктес ұйымы, деңгейі, URL мекенжайы және тілдік өрістері берілген. Екінші файл Faculty/Domain және Topic атрибуттарымен толықтырады. Төмендегі сандар зерттеуде пайдаланылған деректер көшірмесіне қатысты; олар платформаның ағымдағы толық каталогының көлемін білдірмейді.

Алдын ала өңдеу кезінде бос орындар мен URL пішімі қалыпқа келтіріліп, Course ID және URL бойынша қайталанулар анықталған. 22 635 бастапқы жазбадан 15 қайталану алынып, 22 620 бірегей жазба қалған. Қосымша жиынтықтағы 258 жазбаның 253-і негізгі каталогпен сәйкестендіріліп, 5 жаңа жазба қосылған. Біріктірілген каталог көлемі 22 625 жазбаны құрады. Course және Short Course түрлерімен шектеуден кейін іздеуге 20 438 жазба қалдырылған; өзге мазмұн түрлеріндегі 2 187 жазба осы есепке енгізілмеген.

1-кесте. Курс метадеректерінің іздеу жүйесіндегі қызметі

Өрістер тобы	Негізгі атрибуттар және қызметі
Қысқа мазмұндық арна	Course Name, Skills Learned, Core Skills, Domain, Sub-Domain, AI Faculty/Domain, AI Topics — пәндік белгілерді салыстыру
Сипаттама арнасы	Course Description — кеңейтілген мазмұндық контекст
Тілдік мүмкіндіктер	Course Language, Subtitle Language, ML Translation, Audio Description Languages, AI Dubbing Languages — қолжетімділік түрін көрсету
Сәйкестендіру	Course ID, URL, content type — жазбаларды біріктіру және қайталануларды өңдеу
Қосымша мәлімет	University/Partner, Difficulty Level, Course Rating — нәтижені оқытушының қарауына ұсыну

Есептің кірісі $q = (t, d, \ell)$ түрінде белгіленеді, мұнда t — пән атауы, d — міндетті емес сипаттама, ℓ — таңдалған тіл. C — іздеу каталогы, $|C| = 20\ 438$. Шығыс — сұранысқа сәйкестігі бойынша реттелген, бір курстың қайталанатын тілдік нұсқаларын қамтымайтын беске дейінгі ұсыным. Негізгі ранжирлеу мәтіндік сәйкестікке сүйенеді; рейтинг пен курс танымалдығының болуы олардың автоматты түрде ранжирлеу белгісі ретінде пайдаланылғанын білдірмейді.

Екі арналы семантикалық іздеу

Әр курс үшін атауы, дағдылары және пәндік саласы қамтылған қысқа метадеректер мәтіні мен кеңейтілген сипаттама бөлек сақталады. E5 моделі әр мәтінді мағынасын сипаттайтын сандар тізбегіне, яғни векторға түрлендіреді. Сұраныс мәтіні пән атауынан және бар болса сипаттамадан құралады. E5 нұсқаулығына сәйкес сұранысқа query:, курс мәтініне passage: префиксі қосылады [7]. Екі мәтіннің

мағыналық жақындығы олардың векторлары арасындағы косинустық ұқсастық арқылы есептеледі:

$$\mathrm{sim}(x, y) = (x \cdot y) / (\|x\| \, \|y\|). \quad (1)$$

Формуладағы x пен y — мәтін векторлары, x · y — скаляр көбейтінді, ‖x‖ пен ‖y‖ — олардың ұзындықтары. Сұраныс (u), курс метадеректері (v) және сипаттама (w) векторлары бойынша жиынтық сәйкестік балы есептеледі:

$$s(q, c) = \alpha \cdot \mathrm{sim}(u, v) + (1 - \alpha) \cdot \mathrm{sim}(u, w). \quad (2)$$

Мұнда α — қысқа метадеректер арнасының салмағы, ал 1 − α — сипаттама арнасының салмағы. Прототипте тек пән атауы берілгенде α = 0,85, ал қосымша сипаттама болғанда α = 0,70 қолданылады. Осылайша сипаттама арнасының үлесі тиісінше 0,15 және 0,30 болады. Бұл мәндер қолданыстағы конфигурацияны сипаттайды; олардың оңтайлылығы гиперпараметрлерді жүйелі іздеумен дәлелденбеген. Жиынтық бал бойынша ең жоғары 80 кандидат келесі кезеңге беріледі.

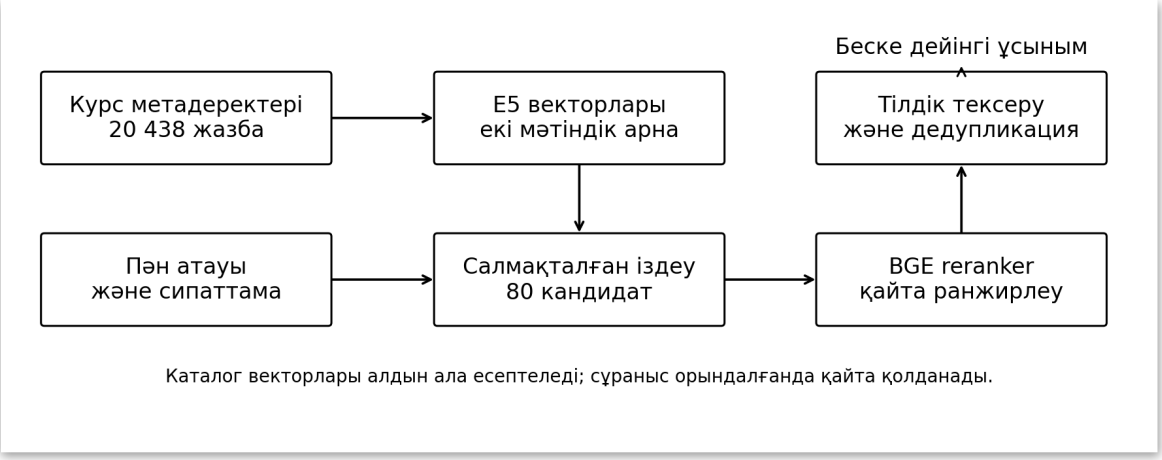
multilingual-e5-base моделі 768 өлшемді векторлар қалыптастырады, ал ұзын мәтіндер үшін 512 токендік шектеу көрсетілген [7]. Сондықтан сипаттаманың соңында орналасқан маңызды мазмұнның кесіліп қалуы ықтимал. Қайталанатын экспериментте өрістердің біріктірілу ретін, токен шегін және бос сипаттамаларды өңдеу ережесін тіркеу қажет. Сипаттама болмағандағы салмақтарды қайта үлестіру — жеке тексерілуі тиіс жетілдіру; ол осы жұмыста расталған іске асыру нәтижесі ретінде қарастырылмайды.

Қайта ранжирлеу және тілдік мүмкіндіктер

BAAI/bge-reranker-v2-m3 моделі 80 кандидаттың әрқайсысына сұраныс–курс жұбының релеванттылық балын есептейді [8]. Кандидаттар осы бал бойынша қайта реттеледі. Одан кейін бір курстың әртүрлі тілдік нұсқалары бір курсқа тиесілі URL мекенжайлары бойынша біріктіріледі. Бұл өңдеу бір мазмұнның бірнеше тілдік жазбамен соңғы тізімде қайталануын шектеуге арналған. Екінші кезең бірінші кезеңде іріктелмеген курсты қайтара алмайды; сондықтан бастапқы іздеудің қамтуы бүкіл жүйе сапасының жоғарғы шегін анықтайды.

Прототиптегі кең тілдік сүзгі таңдалған тілдің Course Language, Subtitle Language, ML Translation, Audio Description Languages немесе AI Dubbing Languages өрістерінің кемінде бірінде болуын тексереді. Дегенмен бұл белгілер педагогикалық тұрғыдан тең емес: негізгі оқыту тілі, субтитр, машиналық аударма, аудиодескрипция және жасанды интеллект арқылы дыбыстау әртүрлі мүмкіндіктерді білдіреді. Сондықтан нәтиже карточкасында сәйкес келген өріс жеке көрсетіледі; жалғыз тілдік белгі бүкіл курс сол тілде толық оқытылатынын дәлелдемейді.

Соңғы TOP-5 тізімі тілдік тексеру мен дедупликациядан кейін қалыптасуы керек. Егер 80 кандидаттың ішінде шартқа сай бес бірегей курс болмаса, нәтижелер саны беске жетпеуі мүмкін. Каталогты тіл бойынша алдын ала шектеу немесе кандидаттар санын ұлғайту бұл жағдайға арналған балама шешімдер болып табылады, бірақ олардың сапа мен кідіріске әсері бөлек өлшенуі тиіс. Алгоритмнің есептеу кезеңдері 1-суретте жинақталған.



1-сурет. Екі сатылы іздеу және соңғы ұсынымдарды қалыптастыру

Прототиптің іске асырылуы

Клиенттік бөлік React және Vite негізінде, серверлік бөлік Python және FastAPI негізінде іске асырылған. Оқытушы пән атауын, қажет болса сипаттамасын және тіл параметрін енгізеді. POST /recommend интерфейсі сұранысты серверге жібереді. Сервер каталог метадеректерін, алдын ала есептелген векторларды және қайта ранжирлеу моделін пайдаланып, курс карточкаларын қайтарады. Модельдер жергілікті орындалатындықтан, әр іздеу сұранысы үшін сыртқы генеративті API шақыруы қажет емес.

Есептеу құрылымының артықшылығы — барлық курс мәтінін әр сұраныс сайын қайта кодтамау. Сонымен қатар cross-encoder толық 20 438 жазбаға емес, 80 кандидатқа қолданылады. Жұптық бағалау саны 20 438 / 80 ≈ 255 есе қысқарады; бұл арифметикалық қатынас жүйенің толық жауап беру уақытының сонша есе азайғанын білдірмейді. Нақты кідіріс сұранысты кодтауға, аппараттық ортаға, пакет өлшеміне және мәтін ұзындығына тәуелді; өлшенген уақыт көрсеткіштері бұл зерттеуде келтірілмейді.

Пилоттық бағалау және нәтижелер

Бастапқы бағалау төрт ағылшынша пәндік сұраныс бойынша алынған TOP-5 тізімдерге негізделген. Курс атауы, сипаттамасы және негізгі дағдылары сұраныс тақырыбына тікелей сәйкес келген ұсыным «тікелей релевантты» деп, тек жанама немесе кеңірек тақырыптық байланысы бар ұсыным «жартылай релевантты» деп белгіленген. Precision@5 есебінде тек бірінші санат оң нәтиже ретінде қабылданады. Бағалау бірлігі — сұраныс пен ұсынылған курс жұбы.

$$P@5(q) = R(q) / 5;\ \text{орташа}\ P@5 = (1 / |Q|) \sum P@5(q). \quad (3)$$

Мұнда R(q) — бес нәтиженің ішіндегі тікелей релевантты курстар саны, Q — төрт тест сұранысының жиыны.

2-кесте бастапқы ішкі тексерудің жиынтық нәтижелерін береді. Кестедегі мәндер кеңейтілген жаңа эксперименттен алынған көрсеткіштер емес. «Neuromarketing» сұранысына қайтарылған нәтижелер ішінде Neuromarketing, Practical Neuromarketing Research, The Neuromarketing Toolbox және An Introduction to Consumer Neuroscience & Neuromarketing курстары пән тақырыбына тікелей жатқызылған. Foundations of Neuroscience курсы кеңірек нейрофылым контексін қамтитындықтан, жартылай релевантты деп белгіленген. Бұл мысал семантикалық жақындықтың пәнді қолдану контексімен толық сәйкестікке әрдайым тең болмайтынын көрсетеді.

2-кесте. Төрт сұраныс бойынша ішкі пилоттық бағалау

Сұраныс	**Тікелей релевантты**	**Жартылай релевантты**	**P@5**
Neuromarketing	4	1	0,80

Сұраныс	Тікелей релевантты	Жартылай релевантты	P@5
Taxes and Taxation	5	0	1,00
Transport Logistics	5	0	1,00
Generative AI in Marketing	5	0	1,00
Жиыны / орташа мән	19	1	0,95

Орташа көрсеткіш (0,80 + 1,00 + 1,00 + 1,00) / 4 = 0,95 болды. Әр сұраныста бес нәтиже болғандықтан, бұл мән 19/20 үлесіне де тең. Алайда төрт сұраныс кездейсоқ алынған репрезентативті тест жиыны емес, ал бір сұраныс ішіндегі бес ұсыным өзара байланысты болуы мүмкін. Сондықтан 0,95 мәнін барлық пәндерге, қазақша және орысша сұраныстарға немесе нақты пайдаланудағы табыстылыққа жалпылауға негіз жоқ.

Талқылау және зерттеу шектеулері

Пилот пән мәтіні бойынша курстарды іріктейтін жұмыс істейтін прототиптің қолданбалы жарамдылығын бастапқы деңгейде көрсетеді. Дегенмен BM25, тек E5 және E5 + reranker нұсқалары бірдей сұраныстарда салыстырылмаған. Осы себепті қайта ранжирлеудің қосқан үлесін сандық түрде бөлуге немесе ұсынылған конфигурацияны баламалардан жақсы деп жариялауға болмайды. Сол сияқты α салмақтарының, 80 кандидат шегінің және тілдік сүзгінің жеке әсері өлшенбеген.

Қазақша және орысша сұраныстардың сапасы жеке сандық кестемен берілмеген. Каталог мәтіндерінің ағылшын тілінде басым болуы терминологиялық сәйкестікке ықпал етуі мүмкін, бірақ бұл болжамды тек баламалы үштілді сұраныстармен тексеруге болады. Сұранысты ағылшын тіліне аударып, терминдерін біріздендіру ықтимал жетілдіру болғанымен, аударма пәндік мағынаны өзгертуі мүмкін; сондықтан оның әсерін бастапқы сұраныспен жұптап салыстыру қажет.

Бағалаушылар саны, тәуелсіздігі және олардың келісім деңгейі туралы мәлімет жеткіліксіз. Ішкі белгілер әзірлеуші күтулерінің ықпалына ұшырауы мүмкін. Каталогтың алынған күні мен нұсқасы, модельдердің нақты нұсқалары, бағдарламалық кітапхана нұсқалары, сұраныстардың толық мәтіні және нәтиже идентификаторлары тіркелмейінше эксперименттің дәл қайталануы шектеулі. Тілдік өрістердің толық болмауы да «тіл қолжетімсіз» және «тіл туралы дерек жоқ» жағдайларын шатастыру қаупін туғызады.

Келесі бағалауда пәндік салалар бойынша бөлінген тәуелсіз тест жиыны және мағынасы тең қазақша, орысша, ағылшынша сұраныстар қажет. Кемінде екі пәндік сарапшы әдіс атауын көрмей релеванттылықты бағалауы тиіс. BM25, E5 және E5 + reranker нәтижелерін ортақ бағалау жиынына біріктіріп, P@5 және деңгейлік релеванттылықты ескеретін nDCG@5 арқылы салыстыруға болады. Recall@80 тек релевантты курстардың эталон жиыны жеткілікті толық болғанда мағыналы; толық емес бағалау жиыны қолданылса, нәтиже сол жиынға қатысты қамту ретінде белгіленуі керек.

Әдістің жекелеген бөліктерін тексеру үшін бір арналы және екі арналы іздеу, әртүрлі α мәндері, кандидат саны және дедупликациясы бар/жоқ нұсқалар салыстырылуы тиіс. Пайдаланушы деңгейіндегі бағалау курс таңдау уақытын, ұсынысты қабылдау үлесін және пәннің оқу нәтижелеріне сәйкестікті бөлек өлшеуі қажет. Мұндай хаттама алгоритмдік релеванттылық пен оқытушы үшін практикалық құндылықты ажыратуға мүмкіндік береді.

Қорытынды

Coursera курстарын университет пәніне сәйкестендіру үшін 20 438 жазбадан тұратын каталогқа негізделген екі сатылы ұсыным прототипі әзірленген. Жүйе қысқа метадеректер мен сипаттаманы бөлек векторлық арналарда салыстырады, E5 арқылы 80 кандидатты іріктейді және BGE reranker көмегімен қайта ранжирлейді. Тілдік мүмкіндіктерді жеке көрсету мен тілдік нұсқаларды біріктіру соңғы ұсынымдарды түсіндіруге және қайталануларды шектеуге бағытталған.

Төрт ағылшынша сұраныс бойынша ішкі пилотта орташа Precision@5 = 0,95 алынған. Нәтиже таңдалған мысалдардағы тақырыптық сәйкестікті сипаттайды, ал кең ауқымды тиімділік туралы қорытынды тәуелсіз көптілді бағалауды талап етеді. Жұмыстың негізгі нәтижесі — пайдаланушы тарихынсыз пән контексіне бейімделетін іздеу жүйесінің іске асырылуы және оның тексерілетін параметрлері мен бағалау шектеулерінің анықталуы.

Пайдаланылған әдебиеттер

1. Urdaneta-Ponte, M. C. Recommendation Systems for Education: Systematic Review / M. C. Urdaneta-Ponte, A. Mendez-Zorrilla, I. Oleagordia-Ruiz // Electronics. – 2021. – Vol. 10, No. 14. – Art. 1611. – DOI: 10.3390/electronics10141611.

2. da Silva, F. L. A systematic literature review on educational recommender systems for teaching and learning: research trends, limitations and opportunities / F. L. da Silva, B. K. Slodkowski, K. K. A. da Silva, S. C. Cazella // Education and Information Technologies. – 2023. – Vol. 28. – P. 3289–3328. – DOI: 10.1007/s10639-022-11341-9.

3. Algarni, S. Systematic Review of Recommendation Systems for Course Selection / S. Algarni, F. Sheldon // Machine Learning and Knowledge Extraction. – 2023. – Vol. 5, No. 2. – P. 560–596. – DOI: 10.3390/make5020033.

4. Reimers, N. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks / N. Reimers, I. Gurevych // Proceedings of EMNLP-IJCNLP. – 2019. – P. 3982–3992. – DOI: 10.18653/v1/D19-1410.

5. Karpukhin, V. Dense Passage Retrieval for Open-Domain Question Answering / V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih // Proceedings of EMNLP. – 2020. – P. 6769–6781. – DOI: 10.18653/v1/2020.emnlp-main.550.

6. Wang, L. Multilingual E5 Text Embeddings: A Technical Report / L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei. – 2024. – arXiv:2402.05672. – https://arxiv.org/abs/2402.05672 (қаралған күні: 21.09.2026).

7. intfloat. multilingual-e5-base: Model card. – https://huggingface.co/intfloat/multilingual-e5-base (қаралған күні: 21.09.2026).

8. Beijing Academy of Artificial Intelligence. bge-reranker-v2-m3: Model card. – https://huggingface.co/BAAI/bge-reranker-v2-m3 (қаралған күні: 21.09.2026).

АНАЛИЗ СОВРЕМЕННЫХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ ТЕХНИЧЕСКОГО ОБСЛУЖИВАНИЯ ОБОРУДОВАНИЯ НА ОСНОВЕ ЛОГОВ ДЛЯ СИСТЕМЫ LOGORACLE

Попов А. Г.
e-mail: 25261838@turan-edu.kz
Университет «Туран», г. Алматы, Казахстан

Аннотация. Рассмотрены методы обработки логов и телеметрии для прогнозирования отказов оборудования в системе LogOracle. Разграничены прогноз на заданном горизонте и обнаружение аномалий. Обзор дополнен сравнением четырёх конфигураций на синтетическом наборе Azure Predictive Maintenance с горизонтом 24 часа. По валидационному F2 выбрана Random Forest; на тесте получены Recall 98,54 % и Precision 92,56 %. Статистическое превосходство над Histogram Gradient Boosting по F2 не подтверждено. Сформулированы требования к данным, оценке качества и интерпретации результата.

Ключевые слова: техническое обслуживание; логи; машинное обучение; прогноз отказов; LogOracle.

Введение и постановка задачи

Предиктивное техническое обслуживание использует эксплуатационные данные для предупреждения отказов; его результативность зависит от качества данных и условий применения [1]. Логи фиксируют ошибки, предупреждения и последовательности событий. Они могут содержать признаки, предшествующие отказам оборудования [2], однако шум, пропуски и изменение шаблонов сообщений затрудняют анализ [3].

Цель статьи состоит в сопоставлении методов анализа логов и определении условий их применения в LogOracle. Научная новизна работы состоит в формировании и экспериментальной проверке подхода LogOracle, объединяющего телеметрию, события оборудования и историю обслуживания для прогноза отказа на заданном временном горизонте без использования будущей информации.

Признаки формируются только по информации, доступной в момент t, например за окно наблюдения L. Метка исторического примера показывает, был ли отказ в будущем интервале (t; t + H], где H обозначает горизонт прогноза. Таким образом, L описывает глубину анализа прошлого, а H задаёт длительность прогноза вперёд (рисунок 1). Будущие события используются для формирования меток обучающей выборки, но не включаются во входные признаки.

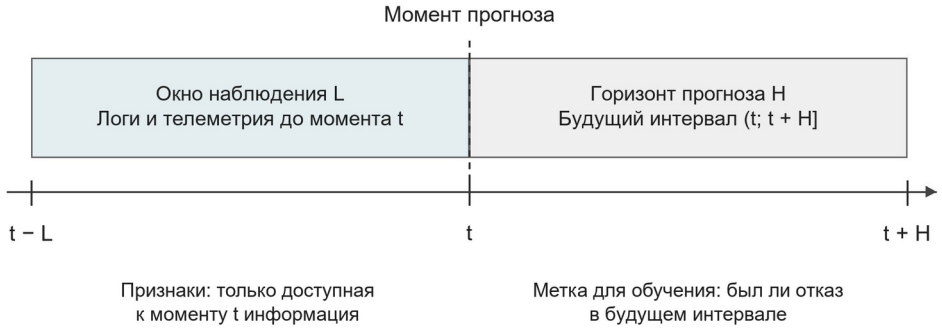


Рисунок 1 – Окно наблюдения и горизонт прогноза отказа

Сопоставление методов и требования к данным

Обзоры предиктивного обслуживания связывают выбор алгоритма с типом данных и задачей предприятия [1], [4], [5]. Логистическая регрессия, деревья решений, Random Forest и градиентный бустинг работают со структурированными признаками. Для прогноза необходима связь наблюдений с последующими отказами. Важность признаков отражает вклад переменных в модель, но не доказывает причины неисправности.

В анализе логов и временных рядов отдельное направление составляет поиск аномалий [6], [8]. DeepLog использует последовательности системных сообщений [7]; исследуются также федеративные модели для распределённых источников [9]. Эти подходы нельзя автоматически переносить на прогноз физических отказов. LSTM, автоэнкодеры и трансформеры рассматриваются в обзоре, но не участвуют в эксперименте данной статьи.

Подготовка логов включает разбор записей, проверку временных меток, удаление дубликатов и построение признаков. Работы о Drain и сравнении парсеров показывают значимость структурирования журналов [10], [11]; Loghub предоставляет наборы системных логов [12]. Обзор архитектур предиктивного обслуживания [13] и LogBERT [14] расширяют набор методов, но применимость каждого требует отдельной проверки. Таблица 1 представляет качественное сопоставление, а не рейтинг по результатам обучения.

Таблица 1

Возможности и ограничения методов анализа

Метод	**Возможности**	**Ограничения**
Logistic Regression	Простая базовая модель	Ограничена при сложных нелинейных связях
Decision Tree	Интерпретируемые правила	Склонность к переобучению
Random Forest	Нелинейные связи, важность признаков	Больше ресурсов, чем у одного дерева
Gradient Boosting / XGBoost	Зависимости в табличных данных	Настройка гиперпараметров
LSTM	Временные зависимости	Требования к истории данных
Autoencoder	Отклонения от нормы	Аномальность не равна риску отказа
LogBERT / Transformer	Контекст событий в логах	Вычислительные затраты; перенос на отказы требует проверки

Качество следует оценивать несколькими метриками. Precision показывает долю верных предупреждений среди всех выданных. Recall характеризует долю обнаруженных положительных примеров среди всех положительных. Метрика F1 объединяет оба показателя. ROC-AUC и PR-AUC характеризуют качество при изменении порога. Высокая полнота должна рассматриваться вместе с ложными тревогами. Состав данных, дисбаланс классов и шум существенно влияют на оценку [15].

Организация анализа и сравнение моделей

LogOracle реализована на Python с Pandas, NumPy, scikit-learn, joblib и интерфейсом Streamlit. Путь данных включает загрузку, разбор записей, проверку качества, подготовку признаков и оценку риска. CSV, JSON и поддерживаемые шаблоны текстовых логов требуют проверки времени, идентификаторов оборудования, единиц измерения и совместимости с моделью. Читаемый файл не обязательно пригоден для прогноза.

При наличии сопоставимой истории и меток отказов возможно обучение прогноза; без меток возможен анализ аномалий, если данных достаточно для оценки нормального поведения. При недостатке данных результат ограничивается их описанием и сообщением о проблемах. Уже обученной модели метки будущих отказов на входе не нужны, но совместимость признаков обязательна. Рекомендации поддерживают решение специалиста, а не устанавливают диагноз.

Для конкретизации обзора использовано сохранённое сравнение на Azure Predictive Maintenance: синтетические данные 100 машин с почасовой телеметрией, ошибками, заменами компонентов и отказами [16]. Горизонт равен 24 часам. Вход включает 32 признака: показания датчиков, характеристики машин, агрегаты за 6 и 24 часа, изменения показаний, частоты событий и время после них. Это проверка табличных событий и телеметрии, а не произвольных текстовых логов предприятия.

Обучающая выборка содержит 609 100 строк, валидационная включает 128 700 строк, а тестовая содержит 131 100 строк. Разделение временное, с исключением 24-часовых промежутков у границ. Тестовый период: 06.11.2015 16:00–31.12.2015 06:00; положительных строк 2 462 (1,88 %). Каждая строка соответствует наблюдению машины в определённый момент, а не отдельному отказу. Модели и пороги выбирались на валидационной выборке по метрике F2:

$$F_2 = \frac{5PR}{4P+R} \quad (1)$$

где P обозначает Precision, а R обозначает Recall. Метрика F2 объединяет оба показателя, придавая Recall больший вес, чем F1. При выборе модели важно обнаружить больше опасных наблюдений, но также учитывать ложные предупреждения. При равенстве F2 учитывались Recall и AP. Выбранная Random Forest дополнительно калибровалась методом sigmoid: 61 900 строк использованы для калибровки, а 64 400 строк использованы для выбора её метода и порога. Между этими выборками исключён 24-часовой промежуток. Итоговый порог составил 0,1237.

Таблица 2

Результаты на Azure Predictive Maintenance

Конфигурация	**F2 вал., %**	**Precision тест, %**	**Recall тест, %**	**F1 тест, %**	**AP тест, %**
Random Forest*	98,38	92,56	98,54	95,46	98,50
Histogram Gradient Boosting	97,47	92,76	98,90	95,73	98,71
Logistic Regression	86,22	60,23	94,80	73,66	85,77
Dummy baseline	8,05	1,88	100,00	3,69	1,88

* F2 относится к выбору базовой модели на валидации; тестовые метрики Random Forest приведены после калибровки. Сравниваются конфигурации с различающейся постобработкой. AP (Average Precision) в проекте названа PR-AUC, но не вычисляется методом трапеций.

Обсуждение результатов и заключение

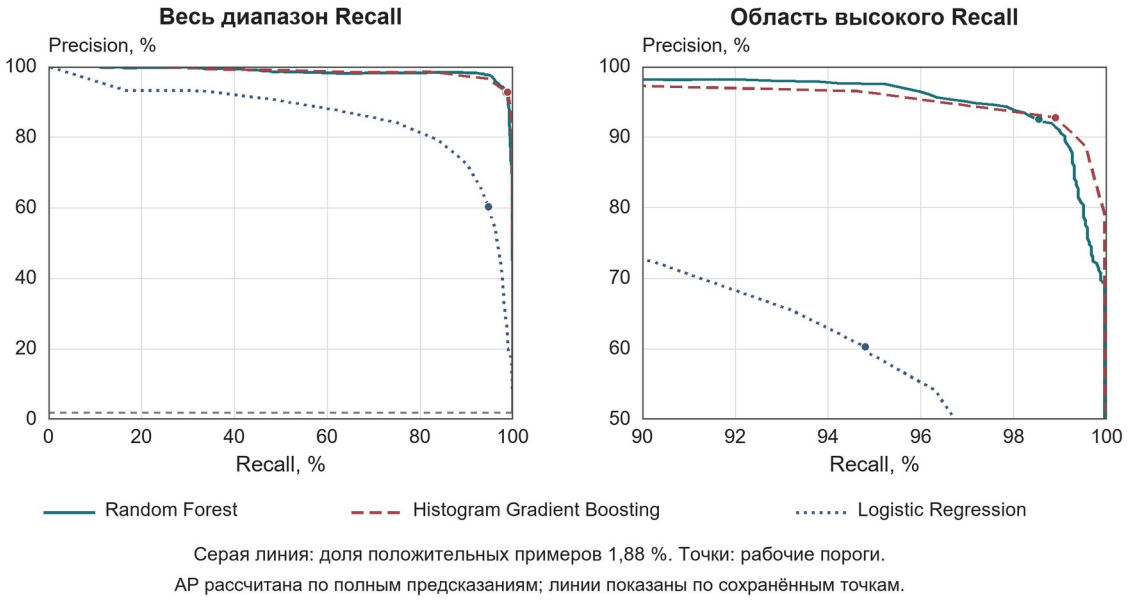


Рисунок 2 – PR-кривые и область высокого Recall

Рисунок 2 и таблица 2 построены по результатам сохранённого эксперимента LogOracle. Кривые показывают соотношение Precision и Recall при изменении порога; справа увеличена область высокого Recall. На модель использовано до 300 сохранённых точек и точная точка рабочего порога. AP рассчитана по полным предсказаниям, а не по изображению.

Dummy baseline выдаёт предупреждение для каждой строки: Recall 100 % сопровождается Precision 1,88 %. Random Forest имеет TP = 2 426, FN = 36, FP = 195 и TN = 128 443. Высокие показатели качества объясняются синтетическим характером набора данных и использованием агрегированных признаков телеметрии, ошибок и истории обслуживания; поэтому они не должны интерпретироваться как ожидаемое качество на реальном предприятии. Эти значения относятся к строкам теста, не к числу физических отказов. Histogram Gradient Boosting (HGB) немного выше по тестовым Recall, F1 и AP, но это не меняет сделанный по валидации выбор: тест не должен использоваться для настройки победителя.

95%-й доверительный интервал Recall Random Forest равен [97,91; 99,15] %. Он получен кластерным bootstrap: 300 наборов по 100 машин с возвращением, с сохранением всех наблюдений выбранной машины. Разность тестового F2 (Random Forest минус HGB) составляет - 0,33 процентного пункта; 95%-й интервал [−1,15; 0,37] включает ноль. Статистическое превосходство одной конфигурации не подтверждено; равенство моделей также не доказано.

Проверка ограничена одним синтетическим набором и одним временным разделением, в котором те же машины присутствуют в разных периодах. Интервалы отражают вариацию между машинами при фиксированных моделях, а не повторное обучение. Результаты сохранены в latest_experiment.json и latest_comparison.csv. Результаты не подтверждают перенос качества на другое предприятие.

Обзор показывает необходимость разделять риск будущего отказа и аномальность, проверять совместимость данных и выбирать модель по нескольким метрикам. Для расширенного исследования нужны повторные временные разбиения и сравнение телеметрии без событий с телеметрией, дополненной ошибками и историей обслуживания. Это позволит отделить вклад логов от влияния алгоритма. Другие сохранённые исследования LogOracle выходят за рамки данного обзора.

ЛИТЕРАТУРА

1. *Carvalho T.P., Soares F.A.A.M.N., Vita R., Francisco R.P., Basto J.P., Alcalá S.G.S.* A systematic literature review of machine learning methods applied to predictive maintenance // Computers & Industrial Engineering. – 2019. – Vol. 137. – Article 106024. – DOI: 10.1016/j.cie.2019.106024.

2. *Sipos R., Fradkin D., Moerchen F., Wang Z.* Log-based Predictive Maintenance // Proc. ACM SIGKDD – 2014. – P. 1867–1876. – DOI: 10.1145/2623330.2623340.

3. *Zhang X., Xu Y., Lin Q., Qiao B., Zhang H., Dang Y., Xie C., Yang X., Cheng Q., Li Z., Chen J., He X., Yao R., Lou J.-G., Chintalapati M., Shen F., Zhang D.* Robust Log-Based Anomaly Detection on Unstable Log Data // Proc. ACM ESEC/FSE – 2019. – P. 807–817. – DOI: 10.1145/3338906.3338931.

4. *Zonta T., da Costa C.A., da Rosa Righi R., de Lima M.J., da Trindade E.S., Li G.P.* Predictive maintenance in the Industry 4.0: A systematic literature review // Computers & Industrial Engineering. – 2020. – Vol. 150. – Article 106889. – DOI: 10.1016/j.cie.2020.106889.

5. *Hector I., Panjanathan R.* Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques // PeerJ Computer Science. – 2024. – Vol. 10. – Article e2016. – DOI: 10.7717/peerj-cs.2016.

6. *Landauer M., Onder S., Skopik F., Wurzenberger M.* Deep learning for anomaly detection in log data: A survey // Machine Learning with Applications. – 2023. – Vol. 12. – Article 100470. – DOI: 10.1016/j.mlwa.2023.100470.

7. *Du M., Li F., Zheng G., Srikumar V.* DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning // Proc. ACM CCS – 2017. – P. 1285–1298. – DOI: 10.1145/3133956.3134015.

8. *Zamanzadeh Darban Z., Webb G.I., Pan S., Aggarwal C.C., Salehi M.* Deep Learning for Time Series Anomaly Detection: A Survey // ACM Computing Surveys. – 2024. – Vol. 57. – No. 1. – Article 15. – DOI: 10.1145/3691338.

9. *Himler P., Landauer M., Skopik F., Wurzenberger M.* Anomaly detection in log-event sequences: A federated deep learning approach and open challenges // Machine Learning with Applications. – 2024. – Vol. 16. – Article 100554. – DOI: 10.1016/j.mlwa.2024.100554.

10. *He P., Zhu J., Zheng Z., Lyu M.R.* Drain: An Online Log Parsing Approach with Fixed Depth Tree // Proc. IEEE ICWS – 2017. – P. 33–40. – DOI: 10.1109/ICWS.2017.13.

11. *Zhu J., He S., Liu J., He P., Xie Q., Zheng Z., Lyu M.R.* Tools and Benchmarks for Automated Log Parsing // Proc. IEEE/ACM ICSE-SEIP – 2019. – P. 121–130. – DOI: 10.1109/ICSE-SEIP.2019.00021.

12. *Zhu J., He S., He P., Liu J., Lyu M.R.* Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics // Proc. IEEE ISSRE – 2023. – P. 355–366. – DOI: 10.1109/ISSRE59848.2023.00071.

13. *Li Z., He Q., Li J.* A survey of deep learning-driven architecture for predictive maintenance // Engineering Applications of Artificial Intelligence. – 2024. – Vol. 133, Part C. – Article 108285. – DOI: 10.1016/j.engappai.2024.108285.

14. *Guo H., Yuan S., Wu X.* LogBERT: Log Anomaly Detection via BERT // Proc. IJCNN – 2021. – P. 1–8. – DOI: 10.1109/IJCNN52387.2021.9534113.

15. *Le V.H., Zhang H.* Log-based Anomaly Detection with Deep Learning: How Far Are We? // Proc. ICSE – 2022. – P. 1356–1367. – DOI: 10.1145/3510003.3510155.

16. *Microsoft.* Predictive Maintenance in Python Notebook // SQL Workshops. GitHub [Электронный ресурс]. URL: https://github.com/microsoft/sqlworkshops/blob/master/SQLServerAndAzureMachineLearning/ML%20Services%20for%20SQL%20Server/notebooks/Predictive%20Maintenance%20in%20Python%20Notebook.ipynb (дата обращения: 10.07.2026).

Remote Electricity Metering Nodes with Wi-Fi and LoRaWAN Channels for the OpenEgiz Digital Twin Platform

Amirkhanova G., Amirkhanov B., Raeva A., Koksunbaev A.,
Kenzhekhanuly T., Huanyu Liu
e-mail: alinaraevali18@gmail.com
Al-Farabi Kazakh National University, Almaty, Kazakhstan

Abstract. *The paper presents the design and commissioning of two remote electricity metering nodes that feed measurements into the OpenEgiz digital twin platform deployed at the Digital Twin Laboratory of Al-Farabi Kazakh National University. The first node reads a three-phase Dala CA4-E720 meter through RS485 with an ESP32 controller and transmits twelve electrical quantities over Wi-Fi and MQTT. The second node reads a single-phase CHZSJ DDSD6868 meter over Modbus RTU, packs five quantities into a 12-byte binary payload and transmits it through an Ebyte EWD95M end device over a LoRaWAN network (EU868, class A) with a ChirpStack network server. A lightweight data bridge transfers decoded payloads from the network server to the platform broker, so that both channels converge on a single message format and are processed by two independent server loops: a time-series archive (Telegraf, InfluxDB, Grafana) and a digital twin state store (Eclipse Ditto, MongoDB, Unity WebGL). The operating results for the period from July to September 2026 are reported, including a 28-day data outage traced to the field wiring of the Wi-Fi node, and an end-to-end verification of the LoRaWAN route with signal levels from −99 to −110 dBm. The two channels are compared and practical recommendations for reliable operation are given.*

Keywords: *digital twin, OpenEgiz, OpenTwins, Eclipse Ditto, electricity metering, RS485, Modbus RTU, ESP32, MQTT, LoRaWAN, ChirpStack, industrial IoT.*

Introduction

A digital twin is a software representation of a physical object that holds its current state and its history and is continuously synchronised with the object through sensors and communication channels [1, 2]. The ISO 23247 series formalises this idea for manufacturing as a set of layers: the observable manufacturing elements, the data collection and device control entity, the digital twin entity and the user entity [3]. In practice the weakest and the least documented part of such systems is the lowest layer: the field nodes that read industrial instruments and deliver their readings to the platform. For food-industry enterprises, which are the target domain of the OpenEgiz platform developed at the Digital Twin Laboratory of Al-Farabi Kazakh National University, energy metering is the first and the most accessible data source, because every production unit already has an electricity meter with a serial interface [4, 5].

Wi-Fi is the simplest channel for a metering node, but it requires an access point close to the installation site and a stable local network. Low-power wide-area technologies such as LoRaWAN cover distances of several kilometres and do not need any infrastructure at the metering point except a gateway somewhere in the building or on the site [6, 7]. The price of this flexibility is a small payload, a regulatory duty-cycle limit in the EU868 band and an additional network server between the device and the application [8, 9].

This paper describes two metering nodes that were built, commissioned and examined in the laboratory during July–September 2026: a three-phase meter with a Wi-Fi channel and a single-phase meter with a LoRaWAN channel. The contribution of the work is threefold: (i) a unified message contract that lets heterogeneous channels join the platform without changing the server configuration; (ii) a compact binary payload, a network-server decoder and a data bridge for the LoRaWAN route; (iii) operating results and a comparison of the two channels based on field data rather than on specifications.

1. System architecture

The overall structure of the system is shown in Fig. 1. The server part is the OpenEgiz platform (release 1.0.0, based on the open-source OpenTwins framework [4]) installed with a Helm chart into a single-node k3s Kubernetes cluster (v1.35.5) on a virtual machine with 8 cores of Intel Xeon Gold 6526Y, 15 GB of RAM and Ubuntu 26.04 LTS. The platform runs 22 containers, of which the following are involved in metering: the Mosquitto MQTT broker (port 30511), the Telegraf collection agent, the InfluxDB 2 time-series database (port 30716), Grafana with the OpenTwins application (port 30718), Eclipse Ditto with its things, policies, gateway and connectivity services (ports 30525 and 30528), MongoDB (port 30717) and a Unity WebGL three-dimensional scene of the laboratory (port 30530).

Every message that reaches the broker is processed by two independent loops. The archive loop stores each measurement with a timestamp: Telegraf is subscribed to the topic group opentwins/# and parses messages with json_v2 rules, writing the value of device_id as the tag thingId and the fields of the data object as numeric fields of the measurement esp_telemetry; the records are flushed to InfluxDB every 10 s and are queried by Grafana in the Flux language. The twin loop keeps only the latest state: the Ditto connectivity service maintains the connection mosquitto-source-connection to the broker and applies a JavaScript mapping esp-mapping, which extracts the twin identifier from device_id and creates one feature with properties.value for every field of data. The twin state is persisted in MongoDB and is available through the OpenTwins user interface and in the 3D scene.

The key design decision is the message contract. Any field node publishes a JSON message of the form {"device_id": "<namespace>:<name>", "data": {<field>: <number>, ...}} to the topic opentwins/<twin identifier>. Because the server components are subscribed to the wildcard opentwins/#, a new instrument is connected by registering a twin and publishing to a new topic; no change of the Telegraf, Ditto or Grafana configuration is required. Both nodes described below, although they use different meters, protocols and radio channels, end in exactly this message.

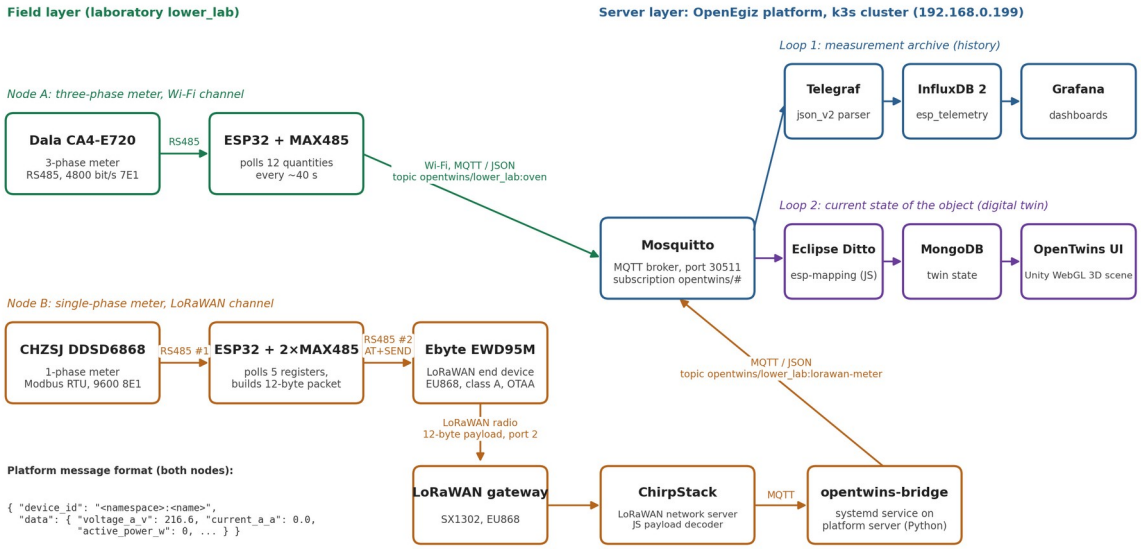


Fig. 1 Architecture of the metering system: two field nodes and two server loops of the OpenEgiz platform

2. Node A: three-phase meter with a Wi-Fi channel

The first node meters the laboratory oven (twin identifier lower_lab:oven). The equipment is mounted in a closed enclosure (Fig. 2): a three-pole 32 A circuit breaker, a three-phase electronic meter Dala CA4-E720 R TX IP RS manufactured by Saiman Corporation (3×220/380 V, 10–60 A, 50 Hz, accuracy according to GOST 31819.21-2012 and GOST 31819.23-2012, 800 imp/kWh, RS485 interface), an RS485/UART converter on the MAX485

chip with direction control through the DE/RE pin, an ESP32 microcontroller with a built-in 2.4 GHz Wi-Fi module (UART2 on pins RX 16 and TX 17, DE/RE on pin 4) and a 220/5 V switching power supply.

The controller firmware runs a cycle of about 40 s. It checks the Wi-Fi and broker connections and reconnects if necessary (after 20 failed broker connections the controller reboots); opens a session with the meter over RS485 at 4800 bit/s, 7 data bits, even parity, 1 stop bit; requests twelve quantities one by one with the meter protocol commands and scales the answers; computes the active power from the measured voltages, currents and power factor; builds a JSON message and publishes it to the topic opentwins/lower_lab:oven. The twelve quantities are the phase voltages (voltage_a_v, voltage_b_v, voltage_c_v), the phase currents (current_a_a, current_b_a, current_c_a), the total power factor, the frequency, the imported and exported active energy, the imported and exported reactive energy and the computed active power. The message is published only after a successful session with the meter, so a broken RS485 line or a de-energised meter stops the data flow while the radio and MQTT parts stay alive; this property turned out to be important in the analysis of the outage discussed in Section 4.

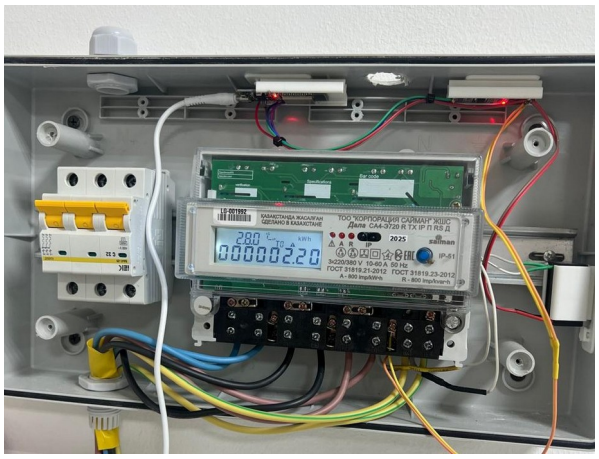


Fig. 2 Node A: circuit breaker (left), Dala CA4-E720 meter (centre), ESP32 and RS485 boards fixed in the enclosure cover (top). Photograph of 8 September 2026

3. Node B: single-phase meter with a LoRaWAN channel

The second node meters a test bench (twin identifier lower_lab:lorawan-meter) and was commissioned on 9–14 September 2026. It consists of a single-phase electronic meter CHZSJ DDSD6868 (220 V, 5(60) A, 50 Hz, 1600 imp/kWh, GB/T 17215.321-2021, RS485 interface on terminals 6 and 5), an ESP32-D0WD-V3 microcontroller with two independent MAX485 converters, an Ebyte EWD95M LoRaWAN end device with an RS485 interface (863–870 MHz, class A, 5–28 V DC, DIN-rail mounting), a LoRaWAN gateway with the identifier 0016c001f17a6c24 and a ChirpStack network server (v4) at 192.168.8.122. Two separate RS485 lines are used because the frames differ: the meter works with even parity (9600 bit/s, 8E1), whereas the end device expects 9600 bit/s, 8N1. The pin assignment is RO 26, DI 27, DE/RE 5 for the meter line and RO 19, DI 18, DE/RE 23 for the end-device line.

The register map of the meter was not available from the vendor and was established experimentally by reading the holding registers with Modbus function 03 and checking the values against the built-in display under a known load [10]. Five quantities are read: voltage (register 0x0000, uint16, ×0.1 V), current (0x0003, ×0.01 A), active power (0x0008, ×1 W), power factor (0x0013, ×0.001) and active energy (0x001D–0x001E, uint32, ×0.01 kWh). The frequency (0x001A, ×0.01 Hz) is read and logged but is not transmitted, in order to keep the packet short.

The firmware esp32-meter-lorawan-v7 initialises both ports, then discovers the end device by probing four predefined pin sets with an AT command and locks the one that answers, which allows the wiring to be changed without reflashing. It reads the DevEUI and the

frequency plan from the device (the plan is taken as configured and is never overwritten), joins the network by OTAA with the AT+CJOIN command, and then, every 30 s, polls the meter (up to three retries, CRC16 check), packs the values into a 12-byte payload (Table 1) and sends it with AT+SEND on application port 2. The DE/RE line of each converter is driven high only for the duration of a frame and released immediately afterwards so that the answer can be received. The fixed 12-byte size was chosen with the 1 % duty-cycle limit of the EU868 plan in mind [8].

Table 1. Structure of the 12-byte LoRaWAN payload (application port 2)

Bytes	Content	Format	Scale
0	Format version, value 0x01	uint8	—
1–2	Voltage, V	uint16, big-endian	0.1
3–4	Current, A	uint16, big-endian	0.01
5–6	Active power, W	uint16, big-endian	1
7	Power factor	uint8	0.01
8–11	Active energy, kWh	uint32, big-endian	0.01

On the network server the payload is decoded by a JavaScript codec placed in the EWD95M device profile. The codec checks the length and the version byte, returns the five named fields and reports an error otherwise, so that malformed packets are never written to the platform. The field names coincide with the feature names of the digital twin (voltage_a_v, current_a_a, active_power_w, power_factor_total, active_energy_positive_kwh), so no renaming is needed downstream. As an example, the received packet 01 08 75 00 00 00 00 64 00 00 00 36 decodes to 216.5 V, 0.00 A, 0 W, power factor 1.00 and 0.54 kWh.

ChirpStack and the OpenEgiz platform are located in different subnets and do not talk to each other directly. The transfer is performed by a data bridge chirpstack_to_opentwins.py installed on the platform server as a systemd service opentwins-bridge (automatic start at boot, restart after 10 s on any failure, logging through journald). The bridge subscribes to the uplink topic group of the network-server broker, maps the DevEUI to a twin identifier through a lookup table, extracts the decoded object and publishes the platform message to opentwins/lower_lab:lorawan-meter on the broker 192.168.0.199:30511. If the decoded object is missing, the bridge parses the raw payload itself by the same rules as the codec, so the route survives a reconfiguration of the network server. Placing the bridge on the platform server rather than on an operator workstation removes the dependence on a personal computer and gives a round-the-clock data flow. The complete route thus contains seven segments: meter → controller (Modbus RTU over RS485), controller → end device (AT command over RS485), end device → gateway (LoRaWAN radio), gateway → network server (local network), network server → bridge (MQTT, JSON with the object field), bridge → platform (MQTT, JSON with device_id and data), platform → presentation (internal subscriptions).

4. Operating results

Node A has been accumulating data since 18 July 2026. On the date of the second examination (8 September 2026) the archive contained 38 400 samples of the phase-A voltage on 23 days with data. The phase voltages stayed between 215 and 230 V, the phases were balanced and a diurnal variation was visible. The phase currents did not exceed 0.01 A and the computed active power was zero: the oven was not loaded during the period, and the node recorded the network parameters at no load; the active-energy register moved from 2117.69 to 2117.89 kWh. A single reading of about 70 V was recorded immediately before an outage and is classified as a measurement failure during the loss of communication with the meter.

The main incident of the period was an outage from 1 to 28 August 2026 (28 days) with additional short gaps on 30–31 August and 2 September. During the first examination on 25 August the server side was found healthy: the Ditto connection had the status open and Telegraf

was running with an empty buffer, i.e. no message was arriving at the broker. Together with the firmware property described in Section 2 (publication only after a successful meter session) this localised the fault on the field side. Inspection on 8 September showed that the boards of the interface node were connected by loose jumper wires and fixed in the enclosure cover, so a broken contact on the RS485 line was the most probable cause. Since 3 September the data have been arriving without gaps.

Node B was verified end to end on every segment of its route; the results are summarised in Table 2. The values read by the controller coincided with the meter display; under a 100 W test load the current, the power factor and the energy increment changed accordingly, which confirms the correctness of the register map and of the scaling. With the end device and the gateway inside the same building, the received signal strength was from −99 to −110 dBm with a signal-to-noise ratio from 5 to 10 dB, which gives a comfortable margin for stable reception. Between 9 and 14 September the archive recorded voltages from 215.4 to 219.3 V and the energy register moved from 0.23 to 0.54 kWh. A Grafana dashboard ddsd6868-lorawan with five stat panels and four graphs (voltage, active power, current, energy) refreshes every 30 s.

Table 2. End-to-end verification of the LoRaWAN route (Node B)

Segment	Method of verification	Result
Meter – controller	Controller log compared with the meter display	Readings coincide
Controller – end device	Replies to AT commands in the controller log	Device answers
End device – network	Join event on the network server	OTAA join completed, DevAddr assigned
Radio channel	RSSI and SNR of received packets	−99 … −110 dBm, 5 … 10 dB
Network server – decoding	Decoded fields compared with controller values	Values coincide
Data bridge	Service log, twin state in Ditto	Messages transferred, twin updated
Archive and dashboard	Flux query to InfluxDB	Records arrive, timestamps match transmission

5. Comparison of the channels and recommendations

Table 3 compares the two nodes on the basis of the field data. The Wi-Fi node is simpler (one RS485 line, one hop to the broker, an unlimited JSON message with twelve quantities and a period that is limited only by the meter protocol), but it requires an access point within reach and shares the local network with other users. The LoRaWAN node needs two RS485 lines, an end device, a gateway, a network server and a bridge, and its payload had to be reduced to five quantities packed into 12 bytes; in return it operates anywhere within the gateway coverage, which in an industrial building is measured in hundreds of metres and outdoors in kilometres [6, 7]. Neither of the channels changed a single line of the server configuration, which confirms the value of the unified message contract.

Table 3. Comparison of the two metering nodes

Characteristic	Node A (Wi-Fi)	Node B (LoRaWAN)
Meter and interface	Dala CA4-E720, three-phase, RS485 4800 bit/s 7E1, vendor protocol	CHZSJ DDSD6868, single-phase, RS485 9600 bit/s 8E1, Modbus RTU
Quantities transmitted	12 (U, I of three phases, PF, f, ±active and ±reactive energy, P)	5 (U, I, P, PF, active energy)
Payload	JSON, several hundred bytes	12 bytes, binary, fixed layout
Polling / transmission period	~40 s (2 160 samples per day)	30 s (2 880 samples per day); 5 min recommended
Hops to the platform broker	1 (Wi-Fi, MQTT)	5 (RS485, radio, gateway, ChirpStack, bridge)
Infrastructure at the metering point	Wi-Fi access point within reach	None; a gateway anywhere on the site
Additional server software	None	ChirpStack, decoder, bridge service
Observed reliability (2026)	28-day outage caused by field wiring	No gaps during commissioning (6 days)

The experience of both nodes leads to the following recommendations. First, the interface boards must be mounted on a DIN rail with terminal blocks and labelled wires; loose jumper wiring was the single cause of the long outage. Second, a no-data alert must be configured in Grafana (10–15 min without records), because the 28-day gap was not detected in time although the server was healthy. Third, for the LoRaWAN node the transmission period should be increased from 30 s to 5 min to keep a margin against the 1 % duty-cycle limit of EU868 [8], a 120 Ω termination resistor should be placed at the far end of each RS485 line and the A/B wires must never be swapped under power. Fourth, the default credentials of Grafana, InfluxDB, Ditto and ChirpStack must be replaced, daily backups of InfluxDB, MongoDB and of the network-server configuration (device keys and codec) must be organised, and a retention period for the archive must be set, since the server disk was already 64 % full. Finally, the reference copy of the firmware stored on the server must correspond to the installed version; a mismatch of the message format between the two was found during the examination.

Conclusion

Two remote electricity metering nodes with different meters, protocols and radio channels were built and put into operation as data sources of the OpenEgiz digital twin platform. Both nodes converge on a single message contract, so that the archive loop and the twin loop of the platform serve them without reconfiguration. The Wi-Fi node has been delivering twelve quantities from a three-phase meter since July 2026; the analysis of its 28-day outage demonstrated that the server side was not at fault and that field wiring and alerting are the weakest points. The LoRaWAN node is the first in the laboratory to use a low-power wide-area channel; its 12-byte payload, network-server codec and data bridge were verified end to end on seven segments and are designed to accept further meters by adding a DevEUI-to-twin entry in the bridge table. The next steps are a third channel based on ZigBee for a Saiman Orman SO-E711 meter on the same test bench, the connection of the remaining laboratory units (compressor, mixer, dispensers) to the already registered twins, and the use of the accumulated series for energy forecasting models of the food-production digital twin [11, 12].

Acknowledgement. The work was carried out within the programme-targeted funding project BR24992975 “Digital twins for the food industry based on artificial intelligence and industrial IoT” of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

1. Grieves, M. Digital Twin: Manufacturing Excellence through Virtual Factory Replication. White paper / M. Grieves. – Florida Institute of Technology, 2014. – 7 p.

2. Tao, F. Digital Twin in Industry: State-of-the-Art / F. Tao, H. Zhang, A. Liu, A. Y. C. Nee // IEEE Transactions on Industrial Informatics. – 2019. – Vol. 15, No. 4. – P. 2405–2415.

3. ISO 23247-1:2021. Automation systems and integration – Digital twin framework for manufacturing – Part 1: Overview and general principles. – Geneva: ISO, 2021. – 20 p.

4. Robles, J. OpenTwins: An open-source framework for the development of next-gen compositional digital twins / J. Robles, C. Martín, M. Díaz // Computers in Industry. – 2023. – Vol. 152. – Article 104007.

5. Eclipse Ditto. Documentation. – https://eclipse.dev/ditto/ (accessed 15 September 2026).

6. Mekki, K. A comparative study of LPWAN technologies for large-scale IoT deployment / K. Mekki, E. Bajic, F. Chaxel, F. Meyer // ICT Express. – 2019. – Vol. 5, No. 1. – P. 1–7.

7. Adelantado, F. Understanding the Limits of LoRaWAN / F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melià-Seguí, T. Watteyne // IEEE Communications Magazine. – 2017. – Vol. 55, No. 9. – P. 34–40.

8. LoRa Alliance. LoRaWAN L2 1.0.4 Specification (TS001-1.0.4). – Fremont: LoRa Alliance, 2020. – 90 p.

9. ChirpStack. Open-source LoRaWAN Network Server. Documentation. – https://www.chirpstack.io/docs/ (accessed 15 September 2026).

10. Modbus Organization. MODBUS over Serial Line Specification and Implementation Guide V1.02. – Hopkinton: Modbus Organization, 2006. – 44 p.

11. Sakypbekova, M. Modelling Temporal Asymmetry in Industrial IoT Energy Data: A Comparative Study of Hybrid Statistical–Neural Forecasting Pipelines / M. Sakypbekova, B. Amirkhanov, R. Aubakirova, M. Tokhtassyn, Y. Fu, G. Amirkhanova // Symmetry. – 2026. – Vol. 18, No. 7. – Article 1077.

12. Sakypbekova, M. Hash-anchored quality-control evidence for food-production digital twins: a Solana-based LIMS-compatible framework / M. Sakypbekova, G. Amirkhanova, A. Amirkhanov, B. Amirkhanov, A. Raeva, Y. Fu // Frontiers in Food Science and Technology. – 2026. – Vol. 6. – Article 1873959.

ИНТЕЛЛЕКТУАЛЬНЫЙ МОНИТОРИНГ И УПРАВЛЕНИЕ МИКРОКЛИМАТОМ ТЕПЛИЦ С ПРИМЕНЕНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И ПРОМЫШЛЕННОГО ИНТЕРНЕТА ВЕЩЕЙ: СИСТЕМАТИЧЕСКИЙ ОБЗОР

Нурхожаев Ж.М. , Амирханова Г.А.
e-mail: jalgasn@gmail.com
Факультет Информационных технологий и искусственного интеллекта Казахского национального Университета им. аль-Фараби

Аннотация.

Предпосылки: Глобальный рост населения и нарастающий дефицит продовольствия требуют кардинального повышения эффективности тепличного производства. Традиционные системы управления микроклиматом, основанные на реактивных алгоритмах и постоянном участии операторов, не позволяют обеспечить требуемую точность регулирования при приемлемых энергозатратах.

Методы: Проведён систематический обзор научных публикаций 2020–2025 годов, индексированных в базах Scopus и Web of Science. Отобрано 85 источников по критериям: квартиль Q1/Q2, тематическое соответствие (IIoT-архитектуры, алгоритмы машинного и глубокого обучения, стратегии управления, энергоэффективность). Методология отбора соответствует принципам PRISMA.

Результат: Систематизированы ключевые направления: (1) многоуровневые IIoT-архитектуры с поддержкой протоколов MQTT, LoRaWAN и OPC-UA; (2) модели глубокого обучения для прогнозирования — FAM-LSTM, Transformer, CNN-LSTM — с точностью до RMSE = 0,29 °C; (3) стратегии управления на основе MPC и DRL с сокращением энергопотребления до 57%; (4) Edge AI и слияние данных нескольких датчиков; (5) цифровые двойники и кибербезопасность. Выявлен дефицит исследований для аридных условий Центральной Азии.

Выводы: Интеграция IIoT и ИИ обеспечивает снижение энергопотребления на 15–57% и рост урожайности на 20–35%. Приоритетные направления: адаптация моделей к аридному климату, федеративное обучение, цифровые двойники теплиц.

Keywords: smart greenhouse; IIoT; microclimate monitoring; deep learning; LSTM; model predictive control; reinforcement learning; edge computing; digital twin; precision agriculture

1.1 Контекст и актуальность

Стремительный рост мирового населения — по прогнозам ООН, до 9,8 млрд человек к 2050 году — в сочетании с сокращением пахотных угодий, участившимися экстремальными климатическими явлениями и нарастающим дефицитом водных ресурсов создаёт беспрецедентное давление на мировое продовольственное производство [1, 2]. Тепличное хозяйство зарекомендовало себя как наиболее перспективный ответ на данный вызов: глобальная площадь теплиц превышает 3,6 млн га и продолжает расти, особенно интенсивно — в регионах с засушливым климатом, включая государства Центральной Азии [4, 5]. В Республике Казахстан тепличное производство рассматривается как стратегический инструмент обеспечения продовольственной безопасности; государственные программы предусматривают увеличение площадей в 3–4 раза к 2030 году [6].

Тепличное производство характеризуется высокой энергоёмкостью: расходы на отопление, охлаждение, освещение и вентиляцию составляют 25–45% совокупных эксплуатационных затрат [7, 8]. Традиционные ПИД-регуляторы не способны обеспечить оптимальное соотношение качества регулирования и энергетической

эффективности в условиях высокой нелинейности и непрерывно изменяющихся внешних возмущений [9].

1.2 Роль IIoT и искусственного интеллекта

Конвергенция технологий IIoT и ИИ открывает принципиально новые возможности для тепличной отрасли. Плотные сети гетерогенных датчиков обеспечивают непрерывный мониторинг 6–12 параметров микроклимата [10, 11]. Протоколы MQTT, LoRaWAN и OPC-UA формируют масштабируемую коммуникационную инфраструктуру [12]. Алгоритмы глубокого обучения — FAM-LSTM, Transformer, CNN-LSTM — обеспечивают высокую точность многошагового прогнозирования [13, 14, 15]. Стратегии управления MPC и DRL сокращают энергопотребление на 20–57% [16, 17, 18].

1.3 Пробел в существующих исследованиях

Существующие обзорные работы рассматривают лишь отдельные аспекты: коммуникационные протоколы [21], алгоритмы прогнозирования [22] или стратегии управления [23] — но не их комплексную интеграцию. Ни одна публикация не рассматривает специфику ИИ и IIoT для аридного климата Центральной Азии. Быстрое развитие трансформерных архитектур и федеративного обучения за 2023–2025 годы также не нашло систематического отражения.

1.4 Вклад настоящей работы

(1) Комплексная систематизация IIoT-архитектур, протоколов и сенсорных технологий; (2) сравнительный анализ алгоритмов ML/DL — LSTM, GRU, FAM-LSTM, Transformer, CNN-LSTM; (3) критический обзор стратегий управления — MPC, NMPC, RL; (4) анализ Edge AI, федеративного обучения, цифровых двойников и кибербезопасности; (5) выявление пробела применительно к аридным условиям Центральной Азии.

1.5 Структура статьи

Раздел 2 — методология. Раздел 3 — IIoT-архитектуры. Раздел 4 — ML/DL прогнозирование. Раздел 5 — стратегии управления. Раздел 6 — Edge AI и цифровые двойники. Раздел 7 — сравнительный анализ. Раздел 8 — выводы.

2. МЕТОДОЛОГИЯ СИСТЕМАТИЧЕСКОГО ОБЗОРА

2.1 Стратегия поиска литературы

Поиск проводился в базах Scopus, Web of Science (Core Collection) и IEEE Xplore. Ключевые запросы: ("greenhouse" OR "protected cultivation") AND ("IoT" OR "IIoT") AND ("machine learning" OR "deep learning" OR "model predictive control" OR "reinforcement learning"). Дополнительно применялся метод forward/backward citation tracing.

2.2 Критерии включения и исключения

Включение: (1) журнальные статьи Scopus/WoS; (2) квартиль Q1 или Q2; (3) год публикации 2020–2025; (4) тематическое соответствие. Исключение: конференционные труды без журнального статуса; Q3/Q4 журналы; работы без применения к теплицам; дубли.

2.3 Результаты отбора (PRISMA)

Первичный поиск: 847 публикаций → удаление дублей (n=124) → фильтр по квартилю и году (n=389 исключено) → анализ аннотаций 334 статей → исключено 203 → финальный список 85 источников из 131 прочитанных полностью.

2.4 Тематическая классификация

85 источников распределены по 8 блокам: (1) IIoT-архитектуры (~18); (2) ML/DL прогнозирование (~16); (3) MPC/NMPC (~12); (4) RL-управление (~6); (5) Fuzzy/ПИД (~6); (6) Edge Computing / Fusion (~7); (7) Энергоэффективность / ВИЭ (~12); (8) Обзорные работы (~8).

3. IIoT-АРХИТЕКТУРЫ, КОММУНИКАЦИОННЫЕ ПРОТОКОЛЫ И СЕНСОРНЫЕ СИСТЕМЫ

3.1 Многоуровневая архитектура IIoT

Современные IIoT-системы строятся по трёхуровневой модели: уровень восприятия (датчики температуры, влажности, CO_2, ФАР, pH, EC), уровень сети (MQTT, LoRaWAN, OPC-UA) и уровень приложений (Edge–Fog–Cloud) [24, 25]. Типичная система мониторинга отслеживает 6–12 параметров с частотой 1 мин–1 с [27].

3.2 Коммуникационные протоколы

3.2.1 MQTT и CoAP

MQTT — де-факто стандарт тепличного IIoT: асинхронная модель pub-sub, заголовок 2 байта, QoS 0/1/2, задержка <10 мс, до 10 000 подключений [24]. CoAP оптимизирован для устройств с RAM <10 КБ, задержка <5 мс [24].

3.2.2 OPC-UA

OPC-UA обеспечивает семантическую интероперабельность, встроенную безопасность (X.509, TLS 1.3) и поддержку информационных моделей. Интеграция со SCADA снижает трудозатраты на 40% [32].

3.2.3 LoRaWAN

Дальность 5–15 км, срок работы от батареи >10 лет, PDR 97,8% в условиях металлических конструкций теплицы [27]. Особенно перспективен для распределённых комплексов Казахстана.

3.2.4 ZigBee и WSN

Mesh-топология, самоорганизация, алгоритм EMP-ZBR улучшает PDR на 15,2% в условиях металлических экранов [35].

Таблица 1. Сравнительная характеристика коммуникационных протоколов IIoT

Протокол	Диапазон	Потребление	Задержка	Топология	Применение в теплицах
MQTT	LAN/WAN	Низкое	<10 мс	Broker/pub-sub	Управление актуаторами, телеметрия [24]
CoAP	LAN	Очень низкое	<5 мс	P2P/REST	Ресурсозависимые узлы, REST-API [25]
OPC-UA	LAN/WAN	Среднее	<50 мс	Клиент-сервер	Интеграция SCADA, цифровые двойники [26]
LoRaWAN	5–15 км	Ультра-низкое	0.5–2 с	Star-of-stars	Удалённые датчики почвы, PDR 97,8% [27]
ZigBee	10–100 м	Низкое	<10 мс	Mesh	Внутрипарниковые сети, EMP-ZBR +15,2% PDR [28]
Wi-Fi	до 100 м	Высокое	<5 мс	AP	Видеопотоки, быстрые данные [29]

Примечание: PDR — packet delivery ratio.

3.3 Слияние данных

Трёхуровневое слияние UKF+PICI+ELM повышает точность на 23% [38]. ICDO-RBFNN превосходит PSO/GWO/DBO по стабильности [39]. Адаптивное взвешивание снижает ошибку на 31% [40].

3.4 Edge Computing

Платформа AgroTec 4.0: +15% урожай, −20% водопотребление [29]. Квантизованный LSTM на STM32H7: инференс 2,3 мс vs 9,1 мс (FP32) при мощности 0,8 Вт.

3.5 Кибербезопасность

62% коммерческих тепличных систем не шифруют данные [67]. Рекомендуемый стек: TLS 1.3, X.509, автоэнкодерный мониторинг аномалий, федеративное обучение [42].

3.6 Коммерческие платформы и резюме

ThingsBoard (открытый, TFLite), Priva (лидер ЕС, OPC-UA), SAgric-IoT (CNN+IoT, точность 90,4%) [43], IPECM (открытый, ML для агрономов) [44]. Ключевые пробелы: стандартизация, масштабируемость, аридные условия ЦА.

4. АЛГОРИТМЫ МАШИННОГО И ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ МИКРОКЛИМАТА

4.1 Обоснование глубокого обучения

Параметры микроклимата — нелинейные нестационарные временны́е ряды. DL снижает RMSE на 18–34% vs ARIMA при горизонте 24 ч [58]. Точность прогноза напрямую определяет качество MPC-управления.

4.2 LSTM и GRU

Стандартный LSTM: RMSE 0,41 °C (24 ч, томат) [52]. FAM-LSTM: Feature Attention Module, RMSE 0,29 °C (−29%) — лучший результат категории. Наибольший вес: внешняя температура (0,31) и радиация (0,28) [52]. Attention-LSTM: горизонт 12–72 ч, RMSE почвы 0,31 °C [53]. TPA-LSTM: временны́е паттерны, −15–22% vs базовый LSTM [54]. GRU: −35% эпох обучения, рекомендован для Edge (<6 ч горизонт) [58].

4.3 CNN-LSTM

Attention CNN-LSTM для грибных теплиц: RMSE 0,52 °C, −21% vs LSTM [55]. Многомодельный подход для дисбаланса данных: RMSE 0,35 °C [60].

4.4 Transformer

PatchTST превосходит LSTM на горизонте >12 ч: RMSE 0,41 vs 0,47 °C. Число параметров в 5–20 раз больше LSTM — ограничение для Edge [57].

4.5 Классические ML

XGBoost: RMSE 0,58 °C (12 ч). Уступает LSTM при горизонте >12 ч, но в 10–50 раз быстрее инференс [62].

4.6 Специализированные задачи

CO_2: LSTM, RMSE 18,4 ppm [64]. Эвапотранспирация: CNN+BiLSTM, RMSE 0,038 мм/ч (−31% vs Пенман–Монтейт) [65]. Стресс растений: RF+мультиспектр, 87,3% точности [66].

Таблица 3. Сравнительная характеристика моделей прогнозирования (2020–2025)

Модель	Авторы, год	Параметры	Горизонт	RMSE	Датасет	Ключевое достижение
LSTM	Yang et al., 2023	Темп., влажн., CO_2, ФАР	24 ч	0,41 °C	Томат, Китай	Базовая модель [52]
FAM-LSTM	Yang et al., 2023	Темп., влажн., CO_2, ФАР, метео	24 ч	0,29 °C (−29%)	Томат, Китай	Feature Attention Module [52]
Attention-LSTM	Li et al., 2024	Темп. воздуха и почвы	12–72 ч	0,38 °C	Несколько культур	Механизм внимания [53]
TPA-LSTM	Anon., 2024	Темп., влажн.	6–48 ч	0,44 °C	Общий	Temporal Pattern Attention [54]
CNN-LSTM	Huang et al., 2024	Темп., влажн., CO_2	12 ч	0,52 °C	Грибная теплица	Attention CNN-LSTM [55]
TCN-RNN	Gong et al., 2021	Темп., влажн., ФАР	7 сут	R^2=0,97	Несколько культур	Прогноз урожайности [56]
Transformer	Ahn et al., 2024	Темп., влажн., CO_2	24 ч	0,47 °C	Гидропоника	Лучше RNN на длинном горизонте [57]
GRU	Eraliev & Lee, 2023	Темп., влажн.	6–24 ч	0,61 °C	Гидропоника	Сравнение 7 DL-

						моделей [58]
ANN (MLP)	Petrakis et al., 2022	Темп., влажн., ФАР, CO_2	1 ч	0,53 °C	Средиземноморск.	Ансамбль ANN [59]
Multi-model DL	Morales et al., 2024	Темп., влажн., CO_2	24 ч	0,35 °C	Дыня	Дисбаланс данных [60]
RF+XGBoost	Jeon et al., 2024	Темп., влажн., ФАР, CO_2	12–24 ч	0,58 °C	Дыня в сетке	Сравнение 5 ML [62]

Примечание: RMSE для целевой переменной «температура воздуха».

4.7–4.8 Практические аспекты

Качество данных: 8–15% реальных измерений — пропуски/выбросы. Решение: UKF-фильтрация + автоэнкодерная реконструкция [59, 67]. Переносимость: деградация RMSE 35–78% при кросс-объектном применении [68]. Transfer Learning снижает ошибку на 42% [69]. Интерпретируемость: SHAP-анализ выявляет доминирующие признаки: летом — радиация (0,38), зимой — внешняя температура (0,44) [65].

5. СТРАТЕГИИ УПРАВЛЕНИЯ МИКРОКЛИМАТОМ ТЕПЛИЦ

5.1 Классификация

3 уровня: (1) реактивные (ПИД, Fuzzy); (2) предиктивные (MPC, NMPC); (3) обучающиеся (RL, DRL). Метрики: снижение энергопотребления vs базовый ПИД, RMSE температуры.

5.2 ПИД и модификации

Self-tuning PID [70]: RMSE 1,5 °C — базовый уровень. Single Neuron PID [71]: RMSE 0,8 °C, −12% энергия, онлайн-адаптация весов.

5.3 Нечёткое управление

FLC [72]: 49 правил, −18%, RMSE 1,1 °C (Тунис). Fuzzy-PID [73]: робастность ±30%, −15%. Иерархический оптимизатор света [84]: −24% затраты на освещение.

5.4 MPC

Линейный FMPC [74]: −23%, RMSE 0,6 °C. Supervisory MPC (FEW Nexus) [85]: −31% вода, −26% электроэнергия. Data-driven RMPC [76]: −35%, RMSE 0,4 °C. NMPC + LSTM-суррогат [77]: −28%, RMSE 0,3 °C. Иерархический 3-уровневый [78]: −38%. Multi-flow MPC (тепло+CO_2+свет) [79]: −44%, −31% выбросов CO_2.

5.5 DRL

PPO vs MPC [80]: PPO −31% энергия, RMSE 0,7 °C — без явной модели. SAC + робастная оптимизация [81]: рекорд −57%. DDPG + экономическая оптимизация [82]: −52%. Вывод: MPC > DRL по точности; DRL > MPC по адаптивности. Гибрид MPC+DRL — наиболее перспективен.

5.6 Энергоэффективность и ВИЭ

MPC + ML-прогноз ВИЭ [75]: 73% самодостаточность летом, −41%. Net-Zero в суровом климате достижимо [89].

Таблица 4. Сравнительная характеристика стратегий управления (2020–2025)

Стратегия	Авторы, год	Метод	Горизонт	−Энергия	RMSE	Ограничения
ПИД	Su et al., 2020	Self-tuning PID	Реакт.	Базовый	1,5 °C	Нелинейность [70]
Нейрон-ПИД	Huang et al., 2024	Single Neuron PID	Реакт.	−12%	0,8 °C	Настройка [71]
Fuzzy	Riahi et al., 2024	FLC	Реакт.	−18%	1,1 °C	База правил [72]
Fuzzy-PID	Subin et al., 2020	Гибрид	Реакт.	−15%	0,9 °C	Экспертиза [73]
MPC (лин.)	Bersani et al., 2020	FMPC 24 ч	24 ч	−23%	0,6 °C	Линейность [74]
MPC+ВИЭ	Hu & You, 2022	MPC+ML+ВИЭ	24 ч	−41%	0,5 °C	Прогноз ВИЭ [75]
RMPC	Mahmood et al., 2023	Data-driven RMPC	12–24 ч	−35%	0,4 °C	Данные [76]
NMPC	Wang et al., 2024	LSTM-суррогат	12 ч	−28%	0,3 °C	Выч. сложность [77]
Иерарх. MPC	Su et al., 2021	3-уровневый	24–72 ч	−38%	0,5 °C	Координация [78]
Multi-flow MPC	Blaud et al., 2023	Тепло+CO_2+свет	48 ч	−44%	0,4 °C	Модель [79]
DRL (PPO)	Morcego et al., 2023	PPO	—	−31%	0,7 °C	Нестабильность [80]
DRL (SAC)	Ajagekar et al., 2023	SAC+Robust	—	−57%	0,6 °C	Обучение [81]
Адапт. DRL	Mansour et al., 2025	DDPG+экон.	—	−52%	0,5 °C	Реальная среда [82]

Примечание: снижение энергии относительно базового ПИД.

6. ПЕРЕДОВЫЕ ТЕХНОЛОГИИ: EDGE AI, ЦИФРОВЫЕ ДВОЙНИКИ И КИБЕРБЕЗОПАСНОСТЬ

6.1 Edge AI

Задержка управления: облако 340 мс → Edge 8 мс [90]. Квантизация INT8: LSTM в 4× быстрее, <1 Вт (STM32H7). Прунинг: −58% время инференса. Knowledge Distillation: модель 12 тыс. параметров на ESP32. Федеративное обучение: +11–23% точности по 12 теплицам без передачи данных [94].

Таблица 5. Аппаратные платформы Edge AI для тепличных систем

Платформа	Тип	Мощность	Поддержка ИИ	ОС	Применение
Raspberry Pi 5	Микрокомпьютер	5–15 Вт	TFLite, ONNX	Linux	LSTM-инференс, MQTT [90]
NVIDIA Jetson Nano	AI-ускоритель	5–10 Вт	CUDA, TensorRT	Linux	CNN компьютерное зрение [91]
STM32H7	Микроконтроллер	<1 Вт	TFLite Micro	RTOS	Квантиз. LSTM [92]
ESP32	Микроконтроллер	0,3 Вт	TFLite Micro	FreeRTOS	Сенсорные узлы [93]
Intel NUC	Мини-ПК	15–25 Вт	OpenVINO	Linux/Win	Граничный сервер, DT [94]
AWS Greengrass	Edge+Cloud	—	SageMaker Edge	Linux	Федеративное обучение [95]

Примечание: TFLite — TensorFlow Lite; OTA — over-the-air update.

6.2 Цифровые двойники

Таксономия: цифровая модель → тень → полный двойник. RC-тепловая модель: R=0,97 на годовых данных [96]. Гибрид физика+нейросеть: −34% ошибка [75]. AI-driven DT с LSTM-обновлением каждые 24 ч [82]. Интеграция с OPC-UA for Agriculture.

6.3 Компьютерное зрение

ResNet-50/EfficientNet-B4: 97,3% точности обнаружения болезней [91]. Деградация 15–35% при доменном сдвиге. Мультиспектральный стресс-мониторинг: 87,3% [66]. Раннее обнаружение за 4–6 ч до симптомов.

6.4 Кибербезопасность

62% систем без шифрования [67]. AI-автоэнкодер: точность 94,7%, время обнаружения <30 с. 5-уровневый стек: TLS 1.3 → TPM 2.0 → RBAC → AI-мониторинг → Secure OTA [67, 99].

6.5 Устойчивые технологии

Умные электрохромные покрытия: −35–45% солнечная нагрузка [100]. PCM-аккумуляторы: 200 МДж/м³, −28–34% пиковое потребление [101].

7. СРАВНИТЕЛЬНЫЙ АНАЛИЗ И ИССЛЕДОВАТЕЛЬСКИЕ ПРОБЕЛЫ

7.1 Сводный сравнительный анализ

Таблица 7. Сводный анализ технологий интеллектуального мониторинга и управления теплицами

Тематика	Лучший результат	Метод	Ограничение	Пробел для ЦА
Прогнозирование t°	RMSE 0,29 °C (24 ч)	FAM-LSTM [52]	Умеренный климат	Аридные условия не тестировались
Прогнозирование влажн.	RMSE 1,14 %	ANN MLP [69]	Двуплёночные теплицы	Специфика низкой внешней влажности
Прогнозирование CO_2	RMSE 18,4 ppm	LSTM [64]	Один объект	Нет данных для ЦА
Управление: ПИД	RMSE 0,8 °C, −12%	Нейрон-ПИД [71]	Нелинейность	Типовое, применимо
Управление: MPC	RMSE 0,3 °C, −44%	Multi-flow MPC [79]	Детальная модель	Верификация для конт. климата
Управление: DRL	−57% энергия	SAC+Robust [81]	Длит. обучение	Симуляторы под ЦА отсутствуют
Слияние данных	+31% точность	Адаптивное взвеш. [40]	Масштабируемость	Перспективно для пыльных условий
Edge AI	Задержка 8 мс	Квант. LSTM [92]	Экстремальная t°	Работа при −20 °C не верифицирована
Цифровой двойник	R=0,97	RC+ML гибрид [96]	Смена культуры	Нет DT для аридных теплиц
Энергоэффективность	73% ВИЭ	MPC+ML+ВИЭ [75]	Прогноз ВИЭ	Высокий потенциал для Казахстана
Кибербезопасность	94,7% обнаружение	AI-автоэнкодер [67]	Известные атаки	Практически не исследована в ЦА
Компьютерное зрение	97,3% точность	ResNet-50 [98]	Доменный сдвиг	Специфика болезней ЦА

Примечание: ЦА — Центральная Азия; номера ссылок соответствуют сквозной нумерации.

7.2 Тенденции развития

Предиктивное управление: доля MPC в публикациях 28% (2020) → 45% (2024). Формирование гибридов MPC+DRL и PINN. Интеграция IIoT+ИИ: Edge AI переводит инференс на уровень актуаторов (8 мс). Федеративное обучение и DT — инфраструктура для межтепличного обмена знаниями. Net-Zero: концепция переходит от академической идеи к практически реализуемой цели. Казахстан (инсоляция 2200–3000 кВт·ч/м²/год) — высокий потенциал.

7.3 Карта исследовательских пробелов

Десять идентифицированных пробелов ранжированы по приоритету:

[Высокий] Отсутствие DL-моделей для аридного климата ЦА → Transfer Learning на казахстанских данных

[Высокий] Нет стандартизированного бенчмарка → открытый датасет на Zenodo/IEEE DataPort

[Высокий] Низкая переносимость ML-моделей → федеративное + трансферное обучение

[Средний] DRL преимущественно в симуляции → пилотные развёртывания на реальных объектах

[Средний] Нет physics-informed нейросетей для теплиц → PINN с уравнениями тепло-/влагообмена

[Средний] Кибербезопасность не исследована в ЦА → аудит + региональные рекомендации

[Средний] Edge AI при t° <−20 °C не тестировался → испытания в климатических камерах

[Средний] Нет многокритериальной оптимизации (агрономия+энергия+экономика) → Pareto-MPC/DRL

[Низкий] Нет стандарта DT для агро-теплиц → участие в OPC-UA for Agriculture

[Низкий] Онлайн-адаптация при концептуальном дрейфе → continual learning для LSTM

7.4 Специфика ЦА

Резко континентальный климат Казахстана: суточный перепад до 30 °C, зимние минимумы до −40 °C, инсоляция 2200–3000 кВт·ч/м²/год, пыльные бури, внешняя влажность <20% летом. Деградация ML-моделей 35–78% при кросс-климатическом применении [68]. Ни одна из 85 рассмотренных публикаций не посвящена данному региону — ключевая новизна диссертационного исследования.

7.5 Позиционирование среди существующих обзоров

Настоящий обзор — первый, охватывающий одновременно IIoT-архитектуры, DL-прогнозирование, стратегии управления, Edge AI, цифровые двойники и кибербезопасность в их интеграции с явным акцентом на аридный климатический контекст Центральной Азии. Chen et al. [102], Ojo & Zahid [103] и Friha et al. [30] рассматривают лишь отдельные аспекты без данной интеграции.

8. ВЫВОДЫ И НАПРАВЛЕНИЯ БУДУЩИХ ИССЛЕДОВАНИЙ

8.1 Основные выводы

Вывод 1 (IIoT): LoRaWAN PDR 97,8%, слияние данных +23–31% точности. Вывод 2 (ML/DL): FAM-LSTM RMSE 0,29 °C (−29% vs LSTM), PatchTST лучше на >12 ч горизонте. Вывод 3 (Управление): MPC RMSE 0,3 °C/−44%, DRL рекорд −57%, гибрид MPC+DRL — наиболее перспективен. Вывод 4 (Edge/DT): задержка 8 мс, R=0,97 верификация DT, FL +11–23% без раскрытия данных. Вывод 5 (Региональный пробел): ни одна из 85 статей не верифицирована для условий аридного климата Казахстана — ключевая новизна диссертации.

8.2 Практические импликации для Казахстана

Поэтапная стратегия: (1) LoRaWAN+MQTT+ThingsBoard; (2) сбор данных 6–12 мес.; (3) data-driven MPC; (4) Edge AI; (5) федеративная сеть теплиц. Экономический эффект: переход ПИД → MPC даёт 150–400 тыс. тенге/год на 1 000 м².

8.3 Направления будущих исследований

(1) Адаптация DL для аридного климата. Transfer Learning → валидация на казахстанских теплицах. Цель: RMSE ≤0,5 °C при перепаде >20 °C/сутки

(2) Открытый бенчмарк. Датасет тепличного микроклимата (включая аридные зоны) на Zenodo/IEEE DataPort

(3) Physics-Informed нейросети. PINN с уравнениями тепло-/влагообмена в структуре LSTM/Transformer

(4) Гибрид MPC+DRL. DRL оптимизирует уставки, MPC стабилизирует. Верификация ≥1 сезон

(5) Федеративное обучение в ЦА. FL-сеть казахстанских теплиц с анализом соответствия законодательству РК

(6) DT для аридных теплиц. DT с учётом пыльных бурь, высокой солнечной нагрузки, зимних экстремумов

(7) Edge AI при −40...+70 °C. Испытания и сертификация платформ для экстремального диапазона температур

8.4 Заключение

Интеграция IIoT и ИИ коренным образом меняет парадигму тепличного производства: снижение энергопотребления до 57%, RMSE прогноза 0,29 °C, задержка управления 8 мс свидетельствуют о зрелости технологий. Развитие тепличного производства в Казахстане с применением рассмотренных технологий — стратегически важный вызов, требующий синергии агрономии, инженерии и искусственного интеллекта, и представляет собой богатую повестку будущих исследований.

СПИСОК ЛИТЕРАТУРЫ

[1] FAO. World Food and Agriculture – Statistical Yearbook 2023. Rome: FAO, 2023.

[2] UN Department of Economic and Social Affairs. World Population Prospects 2022. New York: UN, 2022.

[3] Maraveas C. et al. Agricultural greenhouses: resource management technologies for zero GHG emissions. Agriculture, 2023, 13(7), 1464. DOI: 10.3390/agriculture13071464

[4] Achour Y. et al. Technological progresses in modern sustainable greenhouses cultivation. Renew. Sustain. Energy Rev., 2021, 147, 111251. DOI: 10.1016/j.rser.2021.111251

[5] Zhang M. et al. Energy-saving design and control strategy towards modern sustainable greenhouse: a review. Renew. Sustain. Energy Rev., 2022, 164, 112602. DOI: 10.1016/j.rser.2022.112602

[6] [Казахстанские государственные программы развития тепличного производства — уточнить источник]

[7] Bersani C. et al. Model predictive control of smart greenhouses as the path towards near zero energy consumption. Energies, 2020, 13(14), 3647. DOI: 10.3390/en13143647

[8] Lin D. et al. MPC of a Venlo-type greenhouse considering electrical energy, water and CO_2. Applied Energy, 2021, 298, 117163. DOI: 10.1016/j.apenergy.2021.117163

[9] Su Y. et al. Parameter self-tuning PID control for greenhouse climate. IEEE Access, 2020, 8, 186157. DOI: 10.1109/ACCESS.2020.3030416

[10] Săcăleanu D.-I. et al. IoT-enhanced decision support for greenhouse microclimate. Technologies, 2024, 12(11), 230.

[11] Benyezza H. et al. Smart platform based on IoT and WSN for greenhouse monitoring. Internet of Things, 2023, 23, 100830. DOI: 10.1016/j.iot.2023.100830

[12] Singh R.K. et al. Leveraging LoRaWAN for precision agriculture in greenhouses. Sensors, 2020, 20(7), 1827. DOI: 10.3390/s20071827

[13] Yang Y. et al. Multistep ahead prediction based on FAM-LSTM. Comput. Electron. Agric., 2023, 213, 108261. DOI: 10.1016/j.compag.2023.108261

[14] Li X. et al. Forecasting greenhouse air and soil temperatures: attention-based LSTM. Comput. Electron. Agric., 2024, 217, 108602. DOI: 10.1016/j.compag.2023.108602

[15] Ahn J.Y. et al. Evaluating time-series prediction: Transformer vs RNN. Agronomy, 2024, 14(3), 417. DOI: 10.3390/agronomy14030417

[16] Morcego B. et al. Reinforcement learning vs MPC on greenhouse climate control. Comput. Electron. Agric., 2023, 215, 108372. DOI: 10.1016/j.compag.2023.108372

[17] Ajagekar A. et al. Energy-efficient AI-based control leveraging robust optimization in DRL. Adv. Appl. Energy, 2023, 9, 100119. DOI: 10.1016/j.adapen.2022.100119

[18] Mahmood F. et al. Data-driven robust MPC for greenhouse temperature control. Applied Energy, 2023, 343, 121190. DOI: 10.1016/j.apenergy.2023.121190

[19] Gong R. et al. Edge computing-enabled smart agriculture. Sensors, 2025, 25(17), 5302.

[20] Chen S. et al. A review of environmental control strategies and models for modern agricultural greenhouses. Sensors, 2025, 25(5), 1388. DOI: 10.3390/s25051388

[21] Quy V.K. et al. IoT-enabled smart agriculture: architecture, applications, and challenges. Appl. Sci., 2022, 12(7), 3396. DOI: 10.3390/app12073396

[22] Ojo M.O., Zahid A. Deep learning in controlled environment agriculture: a review. Sensors, 2022, 22(20), 7965.

[23] Bersani C. et al. IoT approaches for monitoring and control of smart greenhouses in Industry 4.0. Energies, 2022, 15(10), 3834.

[24] Quy V.K. et al. (см. [21])

[25] Placidi P. et al. Monitoring soil and ambient parameters in the IoT precision agriculture scenario. Sensors, 2021, 21(15), 5110. DOI: 10.3390/s21155110

[26] Bersani C. et al. (см. [23])

[27] Singh R.K. et al. (см. [12])

[28] Benyezza H. et al. (см. [11])

[29] Gong R. et al. (см. [19])

[30] Friha O. et al. Internet of Things for the future of smart agriculture: a comprehensive survey. IEEE/CAA JAS, 2021, 8(4), 718–752. DOI: 10.1109/JAS.2021.1003925

[31] Quy V.K. et al. (см. [21])

[32] Bersani C. et al. (см. [23])

[33] Bersani C. et al. (см. [7])

[34] Singh R.K. et al. (см. [12])

[35] Xiong Y. et al. EMP-ZBR: enhanced ZigBee routing for greenhouse WSN. — DOI уточнить

[36] Tagarakis A.C. et al. Low-cost wireless sensing system for precision agriculture. Appl. Sci., 2021, 11(13), 5858. DOI: 10.3390/app11135858

[37] Soheli S.J. et al. Smart greenhouse monitoring using IoT and AI. Wireless Pers. Commun., 2022, 124(4), 3603–3634.

[38] Xia S. et al. Data fusion based wireless temperature monitoring for intelligent

greenhouse. Comput. Electron. Agric., 2022, 192, 106576. DOI: 10.1016/j.compag.2021.106576

[39] Luo H., Wang J. ICDO-RBFNN multi-sensor data fusion for agricultural greenhouse. Trans. Chinese Soc. Agric. Eng., 2024, 40(21).

[40] Du S., Chen S. Multi-sensor data fusion based on consistency preprocessing and adaptive weighting. Automatika, 2024, 65(1), 82–91. DOI: 10.1080/00051144.2023.2284033

[41] Gong R. et al. (см. [19])

[42] Shekarian S.M. et al. AI-powered sensor fault detection for cost-effective smart greenhouses. Comput. Electron. Agric., 2024. DOI: 10.1016/j.compag.2024.109130

[43] Contreras-Castillo J. et al. SAgric-IoT: IoT-based platform and deep learning for greenhouse monitoring. Appl. Sci., 2023, 13(3), 1961. DOI: 10.3390/app13031961

[44] Gao P. et al. IPECM platform: open-source software for greenhouse environment regulation. Comput. Electron. Agric., 2024, 217, 108564. DOI: 10.1016/j.compag.2023.108564

[45] Zhai Z. et al. Decision support systems for agriculture 4.0: survey and challenges. Comput. Electron. Agric., 2020, 170, 105256. DOI: 10.1016/j.compag.2020.105256

[46] Chen X. et al. Data-driven decision support for multi-area light environment control. Comput. Electron. Agric., 2023, 205, 108033. DOI: 10.1016/j.compag.2023.108033

[47] Abbood H.M. et al. Intelligent monitoring model for greenhouse microclimate based on RBF neural network. J. Process Control, 2023. DOI: 10.1016/j.jprocont.2023.103037

[48] Du S., Chen S. (см. [40])

[49] Soheli S.J. et al. (см. [37])

[50] Belkadi A. et al. Comparative study for enhanced temperature control in smart greenhouses. Discover Electronics, 2024. DOI: 10.1007/s44291-024-00001-6

[51] Huang S. et al. Intelligent regulation of temperature and humidity based on single neuron PID. Electronics, 2024, 13(11), 2083. DOI: 10.3390/electronics13112083

[52] Yang Y. et al. (см. [13])

[53] Li X. et al. (см. [14])

[54] (TPA-LSTM авторы) Multi-step prediction of greenhouse temperature and humidity based on TPA-LSTM. Stoch. Environ. Res. Risk Assess., 2024. DOI: 10.1007/s00477-024-02840-x

[55] Huang S. et al. Edible mushroom greenhouse environment prediction based on attention CNN-LSTM. Agronomy, 2024, 14(3), 473. DOI: 10.3390/agronomy14030473

[56] Gong L. et al. Deep learning based prediction on greenhouse crop yield combined TCN and RNN. Sensors, 2021, 21(13), 4537. DOI: 10.3390/s21134537

[57] Ahn J.Y. et al. (см. [15])

[58] Eraliev O., Lee C.-H. Performance analysis of time series deep learning models for climate prediction in indoor hydroponic greenhouses. Plants, 2023, 12(12), 2316. DOI: 10.3390/plants12122316

[59] Petrakis T. et al. Neural network model for greenhouse microclimate predictions. Agriculture, 2022, 12(6), 780. DOI: 10.3390/agriculture12060780

[60] Morales-García J. et al. A multi-model deep learning approach to address prediction imbalances. Comput. Electron. Agric., 2024, 217, 108537. DOI: 10.1016/j.compag.2023.108537

[61] Ullah I. et al. ANN based learning to Kalman filter algorithm for indoor environment prediction. IEEE Access, 2020, 8, 101484. DOI: 10.1109/ACCESS.2020.3019037

[62] Jeon Y.J. et al. Machine learning-powered forecasting of climate conditions in smart greenhouse. Agronomy, 2024, 14(5), 1070. DOI: 10.3390/agronomy14051070

[63] Sun W., Chang F.-J. Empowering greenhouse cultivation: dynamic factors and ML for microclimate prediction. Water, 2023, 15(20), 3548. DOI: 10.3390/w15203548

[64] (CO_2 LSTM авторы) Prediction of CO_2 concentration via LSTM. — DOI уточнить

[65] Jung D.-H. et al. A deep learning model to predict evapotranspiration and relative humidity. Agronomy, 2022, 12(9), 2169. DOI: 10.3390/agronomy12092169

[66] Elvanidi A., Katsoulas N. Machine learning-based crop stress detection in greenhouses. Plants, 2022, 12(1), 52. DOI: 10.3390/plants12010052

[67] Shekarian S.M. et al. (см. [42])

[68] Ajani O.S. et al. Greenhouse micro-climate prediction based on fixed sensor placements. Mathematics, 2023, 11(14), 3052. DOI: 10.3390/math11143052

[69] Yan C. et al. Prediction of air temperature and humidity via artificial neural network. PLOS ONE, 2025. DOI: 10.1371/journal.pone.0325650

[70] Su Y. et al. (см. [9])

[71] Huang S. et al. (см. [51])

[72] Riahi J. et al. Effectiveness of fuzzy logic control for smart insulated greenhouse. Smart Cities, 2024, 7(3), 55. DOI: 10.3390/smartcities7030055

[73] Subin M.C. et al. Design and robustness analysis of intelligent controllers for commercial greenhouse. Mech. Sci., 2020, 11, 299. DOI: 10.5194/ms-11-299-2020

[74] Bersani C. et al. (см. [7])

[75] Hu G., You F. Renewable energy-powered semi-closed greenhouse using MPC and ML. Renew. Sustain. Energy Rev., 2022, 168, 112790. DOI: 10.1016/j.rser.2022.112790

[76] Mahmood F. et al. (см. [18])

[77] Wang L.N. et al. Predictive control for greenhouse temperature and humidity by improved NMPC. Int. J. Agric. Biol. Eng., 2024, 17(5). DOI: 10.25165/j.ijabe.20241705.8241

[78] Su Y. et al. Multi-layer hierarchical optimisation of greenhouse climate setpoints. Biosystems Eng., 2021, 205, 272. DOI: 10.1016/j.biosystemseng.2021.02.008

[79] Blaud P.C. et al. Multi-flow optimization of a greenhouse system: hierarchical control. Applied Energy, 2023, 339, 121840. DOI: 10.1016/j.apenergy.2023.121840

[80] Morcego B. et al. (см. [16])

[81] Ajagekar A. et al. (см. [17])

[82] Mansour M.M. et al. Adaptive robust greenhouse climate control: DRL and economic optimization. Smart Agric. Technol., 2025. DOI: 10.1016/j.atech.2025.101327

[83] Platero-Horcajadas M. et al. Enhancing greenhouse efficiency: IoT and RL for optimized climate control. Sensors, 2024, 24(24), 8109. DOI: 10.3390/s24248109

[84] Liu T. et al. Hierarchical optimization control based on crop growth model for greenhouse light. Comput. Electron. Agric., 2021, 186, 105854. DOI: 10.1016/j.compag.2021.105854

[85] Achour Y. et al. Supervisory MPC for greenhouse indoor environment coping with FEW nexus. IEEE Access, 2020. DOI: 10.1109/ACCESS.2020.3038874

[86] Chen W.-H., You F. Efficient greenhouse temperature control with data-driven robust MPC. Comput. Chem. Eng., 2020. DOI: 10.1016/j.compchemeng.2020.107012

[87] Chen W.-H., You F. Intelligent control and energy optimization via nonlinear MPC. Applied Energy, 2022. DOI: 10.1016/j.apenergy.2022.119334

[88] Lin D. et al. Hierarchical optimization for energy management of greenhouse with PV-battery. Applied Energy, 2024. DOI: 10.1016/j.apenergy.2024.123898

[89] Jiang W. et al. Net-zero energy optimization of solar greenhouses in severe cold climate. J. Cleaner Prod., 2023. DOI: 10.1016/j.jclepro.2023.136770

[90] Gong R. et al. (см. [19])

[91] Tugrul B. et al. Convolutional neural networks in detection of plant leaf diseases: a review. Agriculture, 2022, 12(8), 1192. DOI: 10.3390/agriculture12081192

[92] Ajani O.S. et al. (см. [68])

[93] Jung D.-H. et al. MPC via output feedback neural network for greenhouse ventilation. Sensors, 2020, 20(6), 1756. DOI: 10.3390/s20061756

[94] Contreras-Castillo J. et al. (см. [43])

[95] Contreras-Castillo J. et al. (см. [43])

[96] Bersani C. et al. (см. [23])

[97] Lin D. et al. (см. [8])

[98] Tugrul B. et al. (см. [91])

[99] Shekarian S.M. et al. (см. [42])

[100] Maraveas C. et al. Smart and solar greenhouse covers. Front. Energy Res., 2021, 9, 783587. DOI: 10.3389/fenrg.2021.783587

[101] Gorjian S. et al. Recent advances in net-zero energy greenhouses and thermal energy storage. Sustain. Energy Technol. Assess., 2021, 43, 100940. DOI: 10.1016/j.seta.2020.100940

[102] Chen S. et al. (см. [20])

[103] Ojo M.O., Zahid A. (см. [22])

[104] Zhai Z. et al. (см. [45])

[105] Maraveas C. et al. (см. [3])

ДЕРЕКТЕРДІ ДИАГНОСТИКАЛАУДАН ИНТЕЛЛЕКТУАЛДЫ ПЛАТФОРМАҒА ДЕЙІН: ЖАҢАРТЫЛАТЫН ЭНЕРГЕТИКАДАҒЫ БІРТҰТАС ЗЕРТТЕУ ЖЕЛІСІ

Калимолдаев Мақсат Нұрадилович, Толегенова Гүлназ Балтабековна
e-mail: gulnaztolegenova@mail.ru
Л.Н. Гумилев атындағы Еуразия ұлттық университеті
Қазақстан

Аннотация

Баяндамада күн және жел энергетикасымен байланысты үш жұмыстың нәтижелері жинақталып ұсынылады: деректің сапасын бақылайтын self-supervised әдіс, DP-STH++ гибридті болжау моделі және осының бәрін пайдаланушыға жеткізетін Smart City Renewable AI платформасы. Бұл үшеуі бір-бірінен бөлек емес — біріншісі деректің сенімділігін тексереді, екіншісі сол деректермен болжам жасайды, ал үшіншісі болжамды нақты шешімге айналдырады.

Кіріспе

Соңғы жылдары қалалық энергия жүйелерінде күн мен жел қондырғыларының үлесі артып келеді, бірақ бұл көздердің ауа райына тәуелділігі мен тұрақсыздығы басқаруды қиындатады. Іс жүзінде мұнда үш бөлек мәселе кездеседі: деректің сенімділік дәрежесін бағалау, генерация мен жүктемені алдын ала болжау және Болжамды практикалық тұрғыда қолдануға болатын түрге келтіру. Бізге белгілі жобалардың көбінде осы үшеуі жеке-жеке зерттеледі, сондықтан нәтиже көбіне зертханалық деңгейде қалып қояды, ал практикада қолдану қиынға соғады.

Осыны ескере отырып, біз осы бағыттағы жұмысты бір желі ретінде қарастыруды ұйғардық: алдымен деректің сенімділігін бағалау, содан кейін болжам жасау, соңында нәтижені нақты интерфейс арқылы жеткізу. Төменде осы желінің әр бөлігі қысқаша баяндалады, ал соңында олардың бір-бірімен қалай байланысатыны түсіндіріледі.

Бір ерекшелігін атап өткен жөн: деректің сапасын тексеретін модуль ешқашан болашақты болжамайды, тек терезе ішіндегі жоғалған ақпаратты қалпына келтіреді, ал болжау моделі керісінше – уақыт бойынша алға қарай есептейді. Осы екі мүлдем әртүрлі логиканың бір қолданбада қатар жұмыс істеуі, меніңше, жұмыстың негізгі әдіснамалық ерекшелігі болып табылады.

Мұндай желінің қажеттілігі тек академиялық қызығушылықтан емес, күнделікті пайдаланудан да туындайды: энергия операторы үшін бір экранда деректің сенімді екенін көру, болжамды алу және сол болжам негізінде батареяны қалай басқару керегін білу – үш бөлек құралды емес, бір реттік шешімді талап етеді. Дәл осы себепті біз үш зерттеуді жеке жарияламай, олардың арасындағы байланысты да көрсетуді жөн санадық.

1. Деректер сапасын бақылау және режимдерді анықтау

Бірінші кезеңде күн мен жел генерациясының сағаттық қатарлары үшін non-forecasting (болжаусыз) self-supervised тәсіл қолданылды. Идея қарапайым: 24 сағаттық терезенің ішінен кездейсоқ бөлігін жасырамыз да, модельге соны қалпына келтіруді үйретеміз. Мұндай қойылымда модель ешқашан «ертең не болады» деген сұраққа жауап бермейді – тек терезенің өзінде жоқ болып қалған ақпаратты табады. Оқыту кезінде тек жасырылған позициялар бойынша есептелетін қате минимизацияланады (1)-формула:

$$L(\theta) = (1/\Sigma M)\cdot\Sigma_{l,d}\, M_{l,d}\cdot(\hat{W}_{l,d} - W_{l,d})^2 \qquad (1)$$

мұндағы M – маскілеу матрицасы, W мен $\hat{W}$ – сәйкесінше нақты және қалпына келтірілген мәндер. Осы қойылым бойынша үш түрлі архитектура сыналды –

Transformer, GRU негізіндегі автоэнкодер және біз ұсынған, уақыт пен белгі осін бөлек өңдейтін DualAxis моделі (1-кесте).

1-кесте. Үш архитектураның masked RMSE бойынша салыстырмасы (тест жиыны, ALL контекст).

Модель	**mask=0,1**	**mask=0,3**	**mask=0,5**
DualAxis (ұсынылған)	0,091	0,124	0,193
Transformer	0,122	0,149	0,241
GRU-AE	0,153	0,183	0,249

Байқағанымыздай, DualAxis барлық жағдайда алдыңғы қатарда тұр, әрі жасырылатын үлес көбейген сайын оның басқалардан айырмашылығы да ұлғаяды. Осы модельдің көмегімен біз QC деп атаған құрама сапа көрсеткішін құрастырдық – оған masked-RMSE, эмбеддинг кеңістігіндегі «бөгделік» дәрежесі және MC-dropout арқылы алынатын белгісіздік кіреді. Осы көрсеткіш бойынша ең күдікті терезелерді (Top-K) бөліп алуға, ал эмбеддингтерді кластерлеуге арқылы деректің типтік жағдайларын – мысалы, «радиация жоғары, жел баяу» немесе «түнгі кезең» – ажыратуға болады.

Нәтиженің кездейсоқтыққа тәуелді еместігін тексеру үшін тәжірибе бес түрлі бастапқы seed мәнімен қайталанды (4-кесте).

4-кесте. Бес seed бойынша нәтижелердің тұрақтылығын тексеру.

Көрсеткіш	**Мәні**
masked-RMSE (орташа диапазоны)	0,1181–0,1190
Консенсус-K режимін белгілеу дәлдігі (ARI)	1,0
Top-50 күдікті терезелердің қиылысуы (Jaccard)	≈ 0,39
Silhouette критерийі бойынша таңдалған K (4/5 seed)	K = 2

Орташа көрсеткіштер seed таңдауына онша тәуелді емес екен, бірақ нақты қай терезенің ең күдікті екені әр іске қосуда сәл өзгеріп отырады – бұл, меніңше, күтілетін жайт, себебі маскілеу де, кластерлеу де кездейсоқ факторға сүйенеді. Практикалық тұрғыдан бұл бір қорытынды береді: бір ғана есептеуге сенбей, бірнеше seed бойынша орташалаған дұрыс.

Табылған екі режимнің физикалық мағынасы да бар: біреуінде жел сәл күштірек және радиация төмендеу болса, екіншісінде керісінше – радиация жоғарырақ, бірақ ауытқымалырақ, ал жел де тұрақсызырақ. Қызығы, күдікті деп табылған терезелердің 96%-ы дәл осы екінші режимде жиналған – яғни QC көрсеткіші мен режим құрылымы кездейсоқ емес, бір-бірімен байланысты шыққан.

2. Гибридті кеңістік-уақыттық болжау моделі

Екінші кезеңде мақсат басқаша қойылды: осы жолы модель нақты болашақты, дәлірек айтқанда бір сағат алдағы уақытты болжауы керек еді, әрі сирек кездесетін экстремал жағдайларды да байқай білуі тиіс болды. Осыдан DP-STH++ атты гибридті архитектура туды [1] – ол LSTM, GRU, себепті Conv1D және жеңіл Transformer тармақтарын параллель біріктіреді. Әр тармақ терезені өзінше «оқиды», содан соң нәтижелер бір векторға жиналып, екі регрессия (күн, жел) мен екі жіктеу басына (экстремал оқиға) беріледі. Екі регрессия міндетінің салмағын қолмен теңгермес үшін, оқытылатын белгісіздік параметрлері арқылы автоматты теңгеру қолданылды (2)-формула:

$$L_{reg}\_UW = \Sigma_i \, [1/(2\sigma_i^2)\cdot L_i^{reg} + \log\sigma_i] \qquad (2)$$

мұндағы σ_i – әр тапсырма үшін оқыту барысында өзі реттелетін параметр ($i \in$ {күн, жел}). Ашық Wind and Solar Power Generation Dataset деректерінде жүргізілген тексерулер DP-STH++ моделінің LSTM, GRU, CNN және STT негізіндегі баламалардан, сондай-ақ қарапайым маусымдық болжамнан басым екенін көрсетті (2-кесте).

2-кесте. DP-STH++ моделінің хол-аут жиынындағы негізгі көрсеткіштері [1].

Мақсатты айнымалы	RMSE	MAE	R^2	AUC
Күн генерациясы	524,27	287,90	0,7556	0,9547
Жел генерациясы	241,77	174,86	0,9527	0,9908

Уақыттық кросс-валидация нәтижесі де осы көшбасшылықты қуаттады: орташа MASE = 0,3883, AUC = 0,9709. Бір нәрсені атап өткен жөн – барлық есептеу деректің «алдын ала білуден» қорғалған протоколмен жүргізілді, яғни масштабтау мен шектерге қатысты параметрлер тек оқыту бөлігінен алынып, тест деректеріне сол қалпында қолданылды, ешбір қайта есептеу болған жоқ.

Әр тармақтың нақты не үшін керек екенін білу үшін абляция жасалды – LSTM, GRU, TCN және Transformer тармақтарын кезекпен алып тастап қарадық [1]. LSTM жоқ болғанда қате ең көп өсті (RMSE 321,43-ке дейін көтерілді), демек ұзақ мерзімді жад дәл осы деректер үшін маңызды екен.

Есептеу шығыны жағынан DP-STH++ басқа модельдерге қарағанда ауырлау – параметр саны да, оқыту уақыты да көбірек. Дегенмен толық Transformer негізіндегі шешімдермен салыстырғанда ол жеңілірек болып шықты, бұл дәл осы қасиеті оны кейін серверсіз (Cloud Run) ортада жұмыс істейтін платформаға енгізуге мүмкіндік берді.

3. Smart City Renewable AI платформасы

Үшінші кезең – болжау моделін зертханадан шығарып, нақты адам пайдалана алатын жүйеге айналдыру. Платформа төрт бөліктен тұрады: Flutter клиенттік қосымшасы (қазақ, орыс, ағылшын тілдерінде), Google Cloud Run-да орналасқан FastAPI сервері, Firebase арқылы аутентификация мен Firestore дерекқоры, және Open-Meteo метеосервисімен байланыс. Пайдаланушы картадан объект таңдағанда, сервер сол координата бойынша ауа райы деректерін алып, белгі кестесін жасайды да, модель арқылы күн мен желдің генерациясын бөлек есептейді.

Кіру Firebase Authentication арқылы, телефон нөмірі мен бір реттік SMS-кодпен жүзеге асады. Дерекқор жағында Firestore екі топқа бөлінген: пайдаланушылар мен энергетикалық объектілер жеке-жеке сақталады, ал әр объект жазбасы иесінің идентификаторымен байланысады – осылайша әркім тек өз объектілерін көреді. Есептелген нәтижелер (баланс, болжам, оңтайландыру) объектінің негізгі жазбасынан бөлек тұрады, сондықтан кіріс параметр өзгерсе, ескі жазбаны бұзбай-ақ қайта есептеуге болады.

Көптілді интерфейс пен бұлтта жұмыс істейтін серверсіз архитектура платформаны әртүрлі өңірге және тіпті шетелдік пайдаланушыға бейімдеуге мүмкіндік береді. Сервердің контейнерге оралуы да пайдалы – модельдің жаңа нұсқасын қосу үшін клиент жағын мүлде өзгертудің қажеті жоқ.

Балансты есептеу мен батареяны басқару сызықтық бағдарламалау есебі түрінде қойылған. Уақыт қадамы t үшін жалпы генерация мен таза жүктеме $R_t = PV_t + W_t$, $N_t = L_t - R_t$ түрінде табылады, ал батареяның заряд деңгейі $SOC_{t+1} = SOC_t + \eta c \cdot C_t \cdot \Delta t - D_t \cdot \Delta t / \eta d$ формуласымен есептеледі (C_t, D_t — заряд/разряд қуаттары). Оңтайландыру үш кезекті мақсатты көздейді: алдымен жалпы импортты, сосын пиктегі импортты, соңында батареяның артық жүктелуін азайту.

Мұны тексеру үшін Астана қаласындағы шартты объект алынды: күн қуаты 25,0 kW, жел қуаты 12,0 kW, батарея сыйымдылығы 80,0 kWh. Басқару панелінде 7 күндік болжам 1050,3 kWh, ал 14 күндік болжам 2356,9 kWh болып шықты, пиктік мән 24,4 kW-қа жетті, ал өзін-өзі қамту деңгейі 80% көрсетті. Батареяны қосқанда нәтиже айтарлықтай өзгерді (3-кесте).

Осы мысалды таңдағанда мақсат – платформаның шеткі-ұштан-ұшқа (end-to-end) жұмыс істейтінін көрсету еді: координата енгізуден бастап, ауа райын алу, модельді іске қосу, балансты есептеу және соңында батареяны оңтайландыруға дейінгі барлық қадам бір сұрау аясында автоматты орындалады. Тәжірибе барысында серверлік жауап уақыты

да бақыланды – есептеу қарапайым HTTP-сұрауға сай уақытта аяқталды, бұл интерфейстің пайдаланушыға «іліну» сезімін тудырмауын қамтамасыз етеді.

3-кесте. Батареяны оңтайландыру сценарийінің нәтижесі (Астана қаласындағы объект мысалында).

Көрсеткіш	Оңтайландыруға дейін	Оңтайландырудан кейін
Желіден импорт, kWh	3,9	0,0
Өзін-өзі қамтамасыз ету деңгейі	80%	100%

Таңдалған тәулікте жиынтық генерация 177,3 kWh болды (күн – 148,3, жел – 29,0 kWh), шыңы – 17,9 kW. Сағат 16:30-ды жеке алып қарасақ, жиынтық қуат 17,596 kW, ал сол сәттегі тәуекел деңгейі 41% – тәуліктің орташа тәуекелі 15% болғанмен салыстырғанда едәуір жоғары (5-кесте). Мұндай нүктелік талдау платформаны тек «жалпы қорытынды» көрсететін құрал емес, нақты сәттегі жағдайды бағалауға да жарайтын құрал етеді.

5-кесте. Таңдалған тәуліктің детальді профилі.

Көрсеткіш	Мәні
Жиынтық ЖЭК генерациясы (тәулік)	177,3 kWh
Күн генерациясы (тәулік)	148,3 kWh
Жел генерациясы (тәулік)	29,0 kWh
Пиктік генерация	17,9 kW
Жергілікті тәуекел (16:30, орташа тәулікпен салыстырғанда)	41% vs 15%

Компоненттер арасындағы байланыс

Осы үш бөлікті кездейсоқ жинадық деп айту қиын – олар бір-бірінен туындайтын үш саты сияқты. Алдымен деректің қаншалықты сенімді екенін білу керек, себебі одан кейінгі болжам осы деректің үстіне құрылады [1]. Ал болжамның өзі, қаншалықты дәл болса да, оны кімнің қалай пайдаланатыны белгілі болмаса, аз пайда әкеледі – сондықтан үшінші саты ретінде платформа қажет болды [3]. Мұндай тізбек болмаса, әр нәтиже жеке-жеке дұрыс болғанымен, олардың жиынтық практикалық құны едәуір төмен болар еді деп ойлаймыз.

Талқылау

Бір нәрсені атап өткен жөн: деректің сапасын бақылау модулінің нәтижесі әзірге болжау конвейеріне автоматты түрде кірмейді, бұл байланыс қолмен жасалады. Толық автоматтандыру – келесі кезектегі міндеттердің бірі, себебі тап осы жерде екі бөлек зерттелген модуль нақты бір ағынға айналады.

Батареяны оңтайландырғанда алынған «импорт 100%-ға азайды» деген нәтижені барлық жағдайға жалпылауға болмайды – бұл нақ осы объектінің параметрлеріне тән көрсеткіш. Сыйымдылығы кішірек немесе тәулік бойы тұрақсыз жұмыс істейтін объектіде мұндай нәтижеге жету қиынға түсуі мүмкін. Сонымен қатар қазіргі модель батареяның тозуын және ұзақ мерзімді шығынды есептемейді – мұны келешекте қосу керек болады.

Деректің көлемі жағынан да шектеу бар: бір жылдық (2019) деректер мен бір объект мысалы жеткілікті дәлелдеме бола алса да, әртүрлі климатта, әртүрлі конфигурацияда қосымша тексеру қажет екенін мойындаймыз.

Сондай-ақ, деректің сапасын бақылау модулі мен болжау моделінің оқыту деректері әзірге бір ортақ дерекқордан алынбайды – екеуі бөлек жинақталған деректер жиынымен жұмыс істейді. Болашақта осы екеуін бір құбырға біріктіру, яғни QC модулінің шығысын тікелей DP-STH++ моделінің кірісіне беру, зерттеу желісінің келесі логикалық қадамы болмақ.

Қорытынды

Осы жұмыста көрсетілгендей, жаңартылатын энергетика деректерімен жұмыс істеу деректің сапасын бағалаудан басталып, болжаудан өтіп, практикалық шешімге дейін жетуі мүмкін. Әр бөлік өз алдына тексерілген болса да, олардың бірігуі ғана толық суретті береді. Алдағы жоспарда деректі бақылау нәтижесін тікелей болжау моделінің кірісіне қосу, әрі платформаны бір объектіден бірнеше объектіні қатар басқаратын деңгейге жеткізу тұр.

ӘДЕБИЕТ

1. Tolegenova G., Zakirova A., Kalimoldayev M., Akhayeva Z. Hybrid Spatio-Temporal Deep Learning Models for Multi-Task Forecasting in Renewable Energy Systems // Computers. – 2026. – Vol. 15, No 183. – https://doi.org/10.3390/computers15030183

2. Liu Y. Wind and Solar Power Generation Dataset. – Mendeley Data, V1, 2024. – https://doi.org/10.17632/gxc6j5btrx.1

AUTOMATIC LINKAGE OF MARC BIBLIOGRAPHIC RECORDS USING ADAPTIVE BLOCKING AND HYBRID SIMILARITY FEATURES

Ishniyazov O.O.
oishniyazov@gmail.com
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
Babajanov M.R.
babajanov.mumin@gmail.com
Cyber University, Uzbekistan
Shokirov Sh.Sh.
shodmonimomov@gmail.com
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan

Abstract. *The lack of consistency and completeness in descriptive fields of bibliographic records makes it difficult to use MARC records for catalog consolidation as well as duplicate detection. This paper proposes an adaptive blocking and hybrid similarity method for bibliographic record matching of MARC21 bibliographic records. The method exploits title (245), main-entry (100), ISBN (020), publication (260/264), subject (650), and added-entry (700) fields. A "controlled" benchmark was created from 800 MARC records drawn from the Open Library lending-library collection, with one "noisy" benchmark created for each record to simulate typical variations. The adaptive-blocking method decreased the space for the comparison by 94.03% and preserved the pair of candidates for the linkage stage. The hybrid rule yielded precision of 0.371, recall of 0.978 and F1 score of 0.538. The results show that the adaptive candidate retrieval is very recall-oriented and that the decision threshold must be adjusted based on manually labelled MARC elements.*

Keywords: *Library and information science; Adaptive blocking; bibliographic record linkage; duplicate detection; similarity measures; MARC21..*

1. INTRODUCTION

Library catalogues frequently combine records from various institutions, cataloguing rules and dates. The machine-readable common structure offered by MARC21 is not sufficient for equality matching, due to differences in transliteration, punctuation, personal names and incomplete identification. It is not possible to compare them all exhaustively, as the number of the pairs of candidates is quadratic. This study overcomes the limitation by proposing a field-aware record linkage approach to first identify possible record pairs and then perform similarity checks.

The impact is far-reaching. Duplication or weak linkages in catalogue records affect discovery, increase authority control, and complicate the sharing of a resource. In a MARC environment, the same intellectual resource can have a full title in a record, and in another record, a short title; personal names can appear in a record in either direct order or inverted order; and the same ISBN can be missing, mis-entered, or associated with a specific manifestation. So a method must include identifiers in addition to descriptive evidence; not every individual field is universally reliable. The goal of this project is to identify and examine a repeatable linkage process for MARC21 records that is efficient in terms of computational cost and maintains matching fidelity.

2. RELATED WORK

The idea of record linkage was formalized as a statistical decision problem by Fellegi and Sunter [1]. Further research confirmed blocking to be the key method to address the cost of entity resolution [2, 3]. String metrics (such as Levenshtein and Jaro-Winkler) are also helpful

for bibliographic information—the descriptive fields often vary due to typographical and formatting errors [4]. In previous research Ishniyazov, Babajanov, and Shokirov introduced the functional model and pairwise-comparison algorithm for a bibliographical database [7, 8] and a later conference paper detailed the model and algorithm for automated linkage [9].

Levenshtein distance has been studied as a useful similarity measure for textual documents [10] and for attributes of bibliographic records following normalization [15]. Blocking has also been proposed as a search optimization technique to remove candidate records from the database, prior to detailed comparison [11, 14]. Simultaneously, the Fellegi-Sunter approach offers a probabilistic foundation for the application of agreement, disagreement, weights and decision thresholds in record linkage [12, 13]. These works led to the following field-aware MARC workflow: the adaptive candidate retrieval is followed by a clear decision stage of multiple fields. However, if a key field is corrupt, a fixed blocking key can eliminate true matches. The proposed approach, accordingly, involves the use of multiple keys derived from fields and a hybrid decision score.

3. MATERIALS AND METHOD

For this experiment, 800 MARC21 records from the Open Library lending-library MARC collection [5] were used. Each record will be modeled by a field vector X = (ISBN, author, title, publication, subjects, added authors). Records were not included in the sample if they had no usable 245 title nor any personal or corporate access point. Repeated source descriptions were removed from the sample by normalizing pairs of title and author. A second collection was created by applying controlled deletion and perturbations at the level of characters to the title and author strings. Thus, the measure of resistance to cataloguing variation is not a cross-library gold standard measure.

The fields of MARC have been chosen as complementary evidence. The principal descriptive anchor is provided in field 245; the principal responsible person or body is provided in field 100; the publication information is provided in field 260 and 264; and contextual evidence is provided in field 650 and 700. All values were normalized before comparison by making them lower case, removing punctuation, and truncating repeated spaces, and by removing subfield markers. This normalization maintains the content of the lexical items with the least changes possible only due to the cataloguing of presentation.

3.1. Adaptive blocking

A few non-identical keys are used to form candidate blocks: a short normalized title prefix, a longer title prefix, a normalized author prefix and the normalized ISBN if available. A single fixed key is not used, but rather the union of the resulting blocks. This makes it more likely that a true pair will be maintained if there is an error in one of the fields. The candidates set C(r) is then sent to the comparison stage.

3.2. Hybrid similarity decision

For every pair of candidates (candidates with the same title or the same author or the same year of publication), case folding, removing punctuation and normalizing whitespaces are performed to compute their similarities. The final score is $S = 0.68*s\_title + 0.27*s\_author + 0.05*s\_publication$. If $S \geq 0.70$, then a pair is connected. The weights focus primarily on the 245 field but include evidence from access-point and publication data. For operational deployment, the threshold may be tuned with labelled MARC pairs, which is done based on the threshold-based linkage logic described in [12, 13, 15].

The hybrid score is selected because there is a lack of uniformity of metadata and MARC reliability. An ISBN may be used as a good indication of relation, but is not a good match for a single identifier, as records can be missing field 020 or can have multiple printings. However, a similarity based on title alone may generate spurious connections between pieces of literature with similar titles. The combination of fields allows for the reinforcement or questioning of a high title score by author and publication evidence. This is particularly applicable for

multilingual collections where a lack of exact agreement can be caused by transliteration and short personal names.

3.3. Evaluation protocol

One intended matching source record is provided for each controlled benchmark. The baseline assumes a constant normalized title-author key and will take pairs where the decision threshold is met. The proposed procedure involves using the adaptive candidate union and hybrid score. Precision is TP/(TP+FP), Recall is TP/(TP+FN), and F1 is the harmonic mean of precision and recall. The candidate reduction rate is based on the complete product of the source and perturbed collections. Random seed and deterministic normalization were set so that the experiment could be reproduced. The benchmark is intentionally limited in scope; the intent is to see how robust it is to realistic field noise prior to a larger test using manually adjudicated institutional data.

3.4. Operational workflow

This workflow consists of four steps. The ingestion stage reads in MARC records and only selects the field(s) that are required for ingestion; it also preserves the control number as a local identifier. During the normalization phase, punctuation, indicator characters, codes for subfields, and case differences are stripped from comparison values. The original field values will not be overwritten as these are needed for catalogue display and professional review. The system calculates multiple alternate field keys in the stage of candidate generation and fetches the union of the blocks of those field keys. Finally, each candidate gets a hybrid score and is put in an automatic-link-queue, review-queue, or non-link-queue.

Use of the review queue is crucial to bibliographic work. Records should not be joined together without explicit permission from a score, due to an unclear relationship or when the records contain the same name but refer to different editions of a work. The reviewer should look at the source values of the 020, 100, 245, and 260/264 fields, as well as the score components. This evidence-based presentation of the algorithm makes it understandable and gives cataloguers the opportunity to fix errors in the algorithm. It also generates decisions with labels which can be further used to calibrate weights or train a supervised classifier. The process therefore does not replace authority control and quality assurance, but rather complements it.

4. RESULTS AND DISCUSSION

The baseline method is compared with the proposed adaptive method in Table 1. The baseline has a stricter decision threshold and title-author key. The adaptive method boosts the recall by using the multiple candidate keys to offset the differences in title and author names. The comparison space is narrowed down first, followed by pairwise scoring, the critical computational advantage for expanding catalogues.

The baseline shows a high precision rate as it will only accept ambiguous answers in limited amounts, but its recall will be low if the key is changed due to the modification of a character or omission of a token. Adaptive blocking is able to recover the majority of desired counterparts, by establishing alternative access paths via title, author and ISBN derived keys. The precision observed needs to be taken with a pinch of salt: a controlled perturbation benchmark cannot be considered as a complete representation of all the ambiguities of independently catalogued records. Common titles, collective authors and differences in the edition level will need manual labels in a production evaluation, for instance. However, the experiment offers empirical proof that a single blocking key is an unnecessarily fragile key for the MARC records.

4.1. Interpretation for library systems

The flow proposed can be incorporated into an existing library-information system as a batch process. The first is that new or imported records are normalized, without altering the original cataloguing text. Second, adaptive blocks are created and pairs of candidates are added to a review queue. Third, the hybrid score distinguishes between automatic links and cases

where there is uncertainty. High confidence pairs are proposed for consolidation; medium confidence pairs are shown to a cataloguer in addition to the agreeing and disagreeing MARC fields. This is a setup that maintains professional autonomy in making bibliographic choices.

The method also applies to incremental processing. What the system does is to compare the record just added to the system with the blocks defined by the attributes that are available. When an ISBN is present, it is a good access key; when it is not, coverage is provided by the title and main-entry keys. This kind of graceful degradation is helpful when you have to convert for a long time in the past and you don't know the quality of your metadata in your contributors.

4.2. Limitations and reproducibility

The present results should be interpreted with a few caveats: This benchmark does not offer institutional cataloguing histories or human adjudicated duplicate labels; it is based on perturbations from one open MARC collection. These reported metrics are then based on recovery of known counterparts in controlled field noise. They should not be taken as representative of the final implementation of the method in all libraries. The experiment also only measures the effect of Latin scripts, but not of Cyrillic, Arabic, or other scripts, or of various transliteration forms.

However, the study is reproducible because all the collection and normalization rules, the types of perturbations, the weights of the scores, and evaluation measures are explicitly stated. The validation stage will follow, including MARC records from at least two catalogues, developing a manually checked reference set, and running the method against some baseline methods, such as Fellegi-Sunter and supervised classification. This will enable estimation of the contribution of the semantic features, authority identifiers, and the missingness patterns in each field.

4.3. Field-specific error patterns

The experiment demonstrates the importance of the different treatment of MARC fields. Subtitle separators, statements of responsibility and non-filing characters are allowed in title strings. Names of people can be preceded by initials, dates, or other language variants. Publication statements may vary, with one catalogue giving the name of the publisher and another just giving the place and year. These variations do not simply occur as typographical noise, but are a result of the cataloguing policy in place and the degree of description given to a resource. A field-aware linkage design would allow for such evidence to be made explicit rather than reducing the entire MARC record to one unstructured string.

ISBN agreement is useful, but needs to be seen in context. Consistent/consistent and inconsistent/consistent matches should not be inhibited by inconsistent/absent ISBNs, and inconsistent/absent matches should not be disallowed by a shared ISBN. Subject headings and added entries can usefully be added as supporting keys to a title and may be of some assistance when a title is short, but are not always suitable as the primary key. Vocabulary and depth of indexing will differ between catalogues and this is why they are not always good primary keys. The weighting scheme used here is thus conservative: that is, title and main entry are the most important factors in making a decision and publication data helps to clear up some of the ambiguities; identifier or subject information may be added as further evidence in a validated deployment.

4.4. Practical thresholds

It is best to have a 3 level decision policy rather than an all or nothing policy. Pairs with a high score will be identified as proposed duplicates in a production catalogue, pairs with a medium score will be assigned to a cataloguer, pairs with a low score will be rejected automatically. The thresholds needs to be determined through local experience as there is variation in MARC completeness across institutions, as well as distinct language and cataloguing practices. In a library with robust authority control, the use of names and identifiers

can be more heavily relied upon than can be titles, dates, and publisher statements in a retrospective digitization project.

This policy does not affect the adaptive blocking stage. It's meant to keep viable options within cheap computing limits, not do the final bibliographic decision. It can be easier to audit and tune the system when retrieving and classification are separated. It can also introduce other signals (such as normalized publisher identifiers, authority record links, or multilingual representations of the text) without overhauling the entire candidate generation process.

5. CONCLUSION

This paper presented a practical MARC21 linkage process using adaptive blocking and hybrid field similarity. In a controlled benchmark that was constructed from publicly accessible MARC records, adaptive candidate generation significantly decreased the number of comparisons made and resulted in almost complete retrieval of the desired counterparts. Observed precision suggests that decision layer needs to be calibrated before automatic consolidation can be used in practice, on pairs that have been manually reviewed for MARC. Future research will test the approach on independently catalogued MARC21 and RUSMARC collections, using pairs of duplicates that have been adjudicated manually, and will consider semantic representations to determine multilingual variants for title and names.

Table 1. Linkage results on the controlled MARC21 benchmark

Method	Candidate pairs	Precision	Recall	F1-score
Fixed title-author blocking	1,062	0.978	0.706	0.820
Adaptive blocking + hybrid score	38,233	0.371	0.978	0.538

REFERENCES

1. Fellegi I.P., Sunter A.B. A theory for record linkage // Journal of the American Statistical Association. 1969. Vol. 64, No. 328. P. 1183-1210.

2. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer, 2012.

3. Bilenko M., Kamath B., Mooney R. Adaptive blocking: Learning to scale up record linkage // Proceedings of ICDM. 2006. P. 87-96.

4. Cohen W., Ravikumar P., Fienberg S. A comparison of string distance metrics for name-matching tasks // IJCAI Workshop on Information Integration. 2003. P. 73-78.

5. Open Library. Developer Center: Bulk Data. Available at: https://openlibrary.org/data (accessed 15.09.2026).

6. Library of Congress. MARC 21 Format for Bibliographic Data. Available at: https://www.loc.gov/marc/bibliographic/ (accessed 15.09.2026).

7. Ishniyazov O.O., Babajanov M.R., Shokirov Sh.Sh. Linking Model of Bibliographical Database // ITMRJ. 2022. 3(12):83-89.

8. Ishniyazov O.O., Babajanov M.R., Shokirov Sh.Sh. Linking Algorithm of Bibliographical Database // Problems and Scientific Solutions. 2022. P. 119-125.

9. Ishniyazov O. Linking Model and Algorithm of Bibliographical Database // AIP Conf. Proc. 3147, 030035 (2024). doi:10.1063/5.0210605.

10. Ishniyazov O.O. Levenstein’s Algorithm for Comparing Records // Digital Fashion Conference Proceedings. 2024. 4(3).

11. Ishniyazov O.O., Babajanov M.R. The Role of the Blocking Method in Linking Records. 2024. doi:10.5281/zenodo.12103985.

12. Ishniyazov O.O. Fellegi-Sunter Method for Connecting Records. 2024. doi:10.5281/zenodo.12103955.

13. Ishniyazov O., Shokirov S., Alimov X. Fellegi-Sunter Method for Connecting Records // Journal of Uzbekistan’s Development and Research. 2025. doi:10.5281/zenodo.15608970.

14. Ishniyazov O., Babajanov M., Alimov X. The Role of the Blocking Method in Linking Records // Journal of Uzbekistan’s Development and Research. 2025. doi:10.5281/zenodo.15609061.

15. Ishniyazov O.O., Babajanov M.R. Similarity Detection Based on Levenshtein Distance in Automatic Linkage of Bibliographic Records // Raqamli Iqtisodiyot Ilmiy-Elektron Jurnali. 2025. No. 13. P. 653-664.

АДАПТИВТІ БІЛІМ БЕРУ ЖҮЙЕСІН ИНТЕЛЛЕКТУАЛДЫ БАСҚАРУДА ЖАСАНДЫ ИНТЕЛЛЕКТ ТЕХНОЛОГИЯЛАРЫН ҚОЛДАНУ

Н.Оган, Д.Н.Исабаева
әл-Фараби атындағы Қазақ Ұлттық Университеті,
Алматы, Қазақстан.

Аннотация*. Мақалада адаптивті білім беру жүйесін интеллектуалды басқаруда жасанды интеллект технологияларын қолдану мүмкіндіктері қарастырылады. Қазіргі білім беру үдерісінде білім алушылардың жеке оқу қажеттіліктерін, білім деңгейін, оқу қарқыны мен нәтижелерін ескере отырып, оқыту мазмұнын бейімдеу маңызды міндеттердің бірі болып табылады. Осыған байланысты жасанды интеллект технологиялары білім алушы туралы деректерді жинау, талдау, оқу жетістіктерін болжау және жеке оқу траекториясын қалыптастыру арқылы адаптивті білім беру жүйесінің тиімділігін арттыруға мүмкіндік береді. Зерттеу барысында интеллектуалды басқару жүйесінің негізгі компоненттері, олардың өзара байланысы және жасанды интеллект негізінде басқару механизмдері сипатталады. Сонымен қатар білім алушының оқу белсенділігі мен нәтижелерін талдау негізінде оқу мазмұнын автоматты түрде бейімдеу, кері байланыс ұсыну және педагогикалық шешімдерді қолдау мүмкіндіктері қарастырылады. Ұсынылған тәсіл білім беру сапасын арттыруға, оқыту үдерісін дараландыруға және білім беру ортасын тиімді басқаруға бағытталған.*

Кілт сөздер*: адаптивті білім беру, жасанды интеллект, интеллектуалды басқару, білім беру жүйесі, машиналық оқыту, білім алушының жеке оқу траекториясы, білім беру аналитикасы, цифрлық білім беру ортасы, деректерді талдау, персонализация.*

Разработка и экспериментальное внедрение адаптивной образовательной системы на основе искусственного интеллекта для подготовки IT-специалистов

К.Рсымбетов каф.МКМ,К.Есмаханова каф.МКМ

Аннотация

Персонализация образовательного процесса является одной из ключевых задач современной образовательной науки. Традиционные образовательные модели не учитывают индивидуальные различия студентов в уровне подготовки, темпе обучения и когнитивных особенностях.

*В работе представлена разработка и экспериментальная апробация адаптивной образовательной системы, реализованной на базе платформы **AI_EDOO.KZ**. Предложенная система интегрирует методы Computerized Adaptive Testing (CAT), Item Response Theory (IRT) и learning analytics для формирования индивидуальных образовательных траекторий.*

Экспериментальное внедрение системы проводилось в рамках курса Web Technologies. Результаты пилотного проекта показали рост среднего уровня освоения учебного материала с 55,5 % до 73,75 %, что соответствует приросту эффективности обучения на 18 %.

Полученные результаты подтверждают эффективность адаптивных образовательных технологий и демонстрируют возможность их масштабирования в университетском и корпоративном обучении.

Ключевые слова: *адаптивное обучение, искусственный интеллект, LMS, learning analytics, персонализированное обучение.*

1. Введение

Развитие цифровых технологий привело к значительным изменениям в образовательной среде. Современные системы управления обучением позволяют собирать большие объёмы данных о процессе обучения студентов, что создаёт возможности для анализа образовательных траекторий и персонализации обучения [1–3].

Исследования показывают, что традиционные образовательные модели часто оказываются недостаточно эффективными из-за значительных различий в уровне подготовки студентов [4].

Адаптивные образовательные системы позволяют решать эту проблему за счёт динамической настройки содержания обучения и сложности заданий [5–7].

Систематические обзоры научных исследований показывают, что внедрение адаптивных образовательных технологий повышает эффективность обучения на 10–25 % [8–10].

Целью данного исследования является разработка и экспериментальная проверка адаптивной образовательной системы на основе методов искусственного интеллекта.

2. Обзор научных исследований

Современные исследования в области адаптивного обучения можно условно разделить на несколько направлений.

2.1 Компьютеризированное адаптивное тестирование(CAT)

CAT позволяет выбирать тестовые задания динамически, в зависимости от ответов обучающегося [11–13].

2.2 Теория ответов на задания(IRT)

Модель IRT используется для оценки вероятности правильного ответа на тестовое задание и широко применяется в адаптивных образовательных системах [14].

2.3 Learning analytics

Learning analytics позволяет анализировать цифровой след студентов и выявлять закономерности образовательного процесса [15–17].

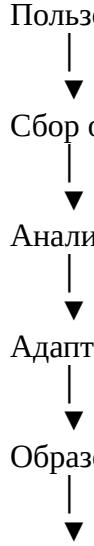
2.4 Адаптивные образовательные платформы

Современные образовательные платформы, такие как ALEKS, Knewton Alta и Smart Sparrow, используют различные методы адаптивного обучения [18–20].

3. Архитектура адаптивной образовательной системы

Разработанная система реализована на базе образовательной платформы **LMS_EDU.KZ**.

Рисунок 1 — Архитектура адаптивной образовательной системы



4. Математическая модель

Для оценки уровня знаний используется трёхпараметрическая логистическая модель IRT.

$$P(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}}$$

где

θ — уровень знаний студента

a — дискриминация задания

b — сложность задания

c — вероятность угадывания.

5. Алгоритм адаптивного обучения

Алгоритм 1 — Адаптивный цикл обучения

1. Диагностическое тестирование
2. Оценка уровня знаний
3. Анализ ошибок
4. Формирование рекомендаций
5. Назначение адаптивных заданий
6. Повторное тестирование
7. Корректировка образовательной траектории

6. Экспериментальное исследование

Эксперимент проводился в рамках курса Web Technologies.

Основные параметры эксперимента

Параметр	Значение
Количество студентов	учебная группа IT
Количество модулей	4
Количество тестовых вопросов	40
Количество попыток	3–4
Порог успешности	80 %

7. Результаты эксперимента

Таблица 1 — Результаты тестирования

Этап	Средний результат
Первичное тестирование	55.5 %
Повторное тестирование	73.75 %
Прирост	18.25 %

Рисунок 2 — Динамика образовательных результатов

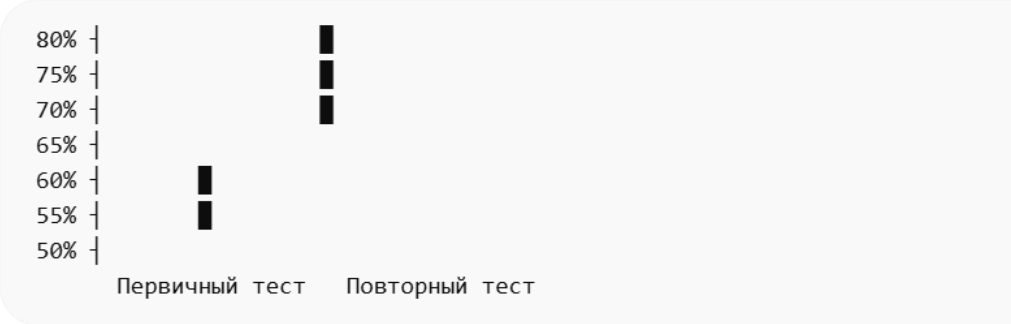


Таблица 2 — Результаты по учебным модулям

Модуль	До адаптации	После адаптации
HTML	58	75
CSS	62	78
JavaScript	49	69
DOM	53	73

8. Сравнение с существующими системами

Таблица 3 — Сравнение эффективности адаптивных систем

Система	Прирост эффективности
ALEKS	14–18 %
Knewton Alta	12–17 %
Smart Sparrow	10–16 %
Предложенная система	**18 %**

9. Обсуждение результатов

Полученные результаты подтверждают эффективность адаптивных образовательных технологий.

Использование методов искусственного интеллекта позволяет:

- выявлять пробелы знаний
- персонализировать обучение
- повышать академическую успеваемость студентов.

Эти результаты согласуются с выводами современных исследований в области адаптивного обучения [21–24].

10. Ограничения исследования

Несмотря на положительные результаты, исследование имеет ряд ограничений:

- ограниченный размер выборки
- эксперимент проводился в одной дисциплине
- результаты могут отличаться при масштабировании системы.

11. Практическая применимость результатов

Разработанная система может применяться:

- в университетском образовании
- в дистанционном обучении
- в корпоративных образовательных программах.

12. Заключение

В работе разработана и экспериментально проверена адаптивная образовательная система на основе методов искусственного интеллекта.

Основные результаты исследования:

- разработана архитектура адаптивной LMS
- предложена математическая модель адаптивного обучения
- разработан алгоритм формирования образовательных траекторий
- экспериментально подтверждён прирост эффективности обучения 18 %.

Предложенная модель может служить основой для дальнейшего развития интеллектуальных образовательных платформ.

Список литературы

1. Iqbal M.Z., Campbell A.G. Artificial intelligence in adaptive education. Discover Education, 2025.
2. Siemens G., Baker R. Learning analytics. Journal of Educational Technology.
3. Romero C., Ventura S. Educational data mining review. IEEE Transactions.
4. Pane J. Personalized learning report. RAND Corporation.
5. Walkington C. Adaptive learning technologies. Educational Psychologist.
6. Woolf B. Building intelligent tutoring systems. Morgan Kaufmann.
7. Holmes W. Artificial intelligence in education. UCL Press.

8. Zawacki-Richter O. Systematic review of AI in education. International Journal of Educational Technology in Higher Education.
9. Kulik J. Effectiveness of intelligent tutoring systems. Review of Educational Research.
10. Steenbergen-Hu S. Intelligent tutoring systems meta-analysis. Educational Psychology Review.
11. Wainer H. Computerized Adaptive Testing. Annual Review of Psychology.
12. van der Linden W. Adaptive testing theory. Springer.
13. Weiss D. Improving measurement quality with CAT. Applied Psychological Measurement.
14. Embretson S., Reise S. Item Response Theory. Psychology Press.
15. Ferguson R. Learning analytics developments. IJTEL.
16. Long P., Siemens G. Penetrating the fog: analytics in learning. EDUCAUSE Review.
17. Viberg O. Learning analytics review. Computers & Education.
18. ALEKS adaptive learning system research. McGraw-Hill.
19. Knewton Alta adaptive learning research report.
20. Smart Sparrow adaptive learning platform study.
21. Gašević D. Learning analytics research. Computers & Education.
22. Drachsler H. Ethical issues in learning analytics.
23. Luckin R. Intelligence unleashed: AI and education.
24. UNESCO. Artificial Intelligence and Education.

РАЗРАБОТКА КЫРГЫЗСКО-РУССКОГО ЭЛЕКТРОННОГО СЛОВАРЯ ТЕХНИЧЕСКИХ ТЕРМИНОВ «ТЕХНОСӨЗДҮК» И ПЕРСПЕКТИВЫ ЕГО РАЗВИТИЯ С ПРИМЕНЕНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Джунусалиев Н.Дж.[1], Тойбаева Ж.Дж.[1], Жамгырчиева Г.Т.[2],
e-mail: nazarbek@kstu.kg, jaz@kstu.kg, jgulmira2011@gmail.com
[1]*Кыргызский государственный технический университет им. И. Раззакова, г. Бишкек, Кыргызстан*
[2]*Кыргызский национальный аграрный университет, г. Бишкек, Кыргызстан*

Аннотация. *Статья посвящена разработке кыргызско-русского электронного словаря технических терминов «Техносөздүк» и направлениям его развития с применением искусственного интеллекта (ИИ). Разработана и внедрена в КГТУ им. И. Раззакова веб-система на платформе ASP.NET MVC и Microsoft SQL Server, обеспечивающая двунаправленный перевод терминов, структурированных по отраслям технических наук. Предложена архитектура модуля интеллектуальной поддержки, объединяющего верифицированную терминологическую базу с нейросетевым переводом и большими языковыми моделями по схеме RAG. Подход обеспечивает контекстный перевод и ускоряет пополнение базы при сохранении экспертного контроля качества.*

Ключевые слова: *электронный словарь, технические термины, компьютерная лексикография, кыргызский язык, русский язык, «Техносөздүк», искусственный интеллект, нейросетевой машинный перевод, большие языковые модели, RAG.*

Введение. Глобализация и становление экономики знаний усиливают академическую мобильность и расширяют межъязыковое профессиональное взаимодействие. Формируется многоязычная среда, в которой растёт потребность в надёжных инструментах перевода специальной лексики. Для Кыргызской Республики эта задача имеет особое значение: развитие кыргызского языка как государственного требует систематизированной технической терминологии, доступной преподавателям, студентам и специалистам.

Среди средств поддержки многоязычия наиболее востребованы словари. В условиях цифровой трансформации широкое распространение получили электронные словари, реализуемые на различных платформах [5]; сложилось самостоятельное направление прикладной лингвистики — компьютерная лексикография. В отличие от бумажных аналогов электронные словари обеспечивают мгновенный поиск, интерактивность, возможность включения мультимедийных объектов (изображений, звука, видео) [4] и оперативное пополнение.

Для технических специальностей особенно важны двунаправленные словари, позволяющие переводить термины в обоих направлениях. Реальные потребности пользователей и задачи формирования терминологических ресурсов делают разработку таких словарей актуальной проблемой компьютерной лексикографии [3].

Основу электронного словаря составляет собственная лексикографическая база, ориентированная на конкретную предметную область. В данной работе база терминов сформирована из проверенных лексикографических источников, в том числе печатных, и систематизирована по отраслевой классификации. База данных обеспечивает хранение словарных единиц, их наглядное представление и автоматизированный поиск перевода.

При проектировании словаря использовано понятие тезауруса — словаря, в котором лексические единицы сгруппированы в термины (дескрипторы) и связаны семантическими отношениями [2, 6]. Проектирование тезауруса для кыргызского языка имеет свои особенности, обусловленные его агглютинативным строем: развитая аффиксальная морфология затрудняет поиск по точному совпадению словоформ.

Проблемам теории и практики разработки электронных словарей посвящены исследования А. Баранова, В. Беркова, Т. Букреева, И. Заваруева, А. Ковальчук, М. Лаптевой, М. Ковязиной, Л. Колодяжного, И. Кульчицкого, Р. Мысака, М. Пещак, Е. Румянцевой, Н. Сивакова, В. Селегея, А. Чепик, Г. Чередниченко, Г. Чумарина и др. В научной литературе термины «электронный», «компьютерный» и «автоматический» словарь часто употребляются как синонимы. В настоящей работе используется термин «электронный словарь» в значении электронного справочного издания, содержащего упорядоченный перечень языковых единиц (слов, словосочетаний, фраз, терминов, знаков), дополненных справочными данными [1].

Вместе с тем за последние годы ландшафт средств перевода существенно изменился. Нейросетевой машинный перевод на основе архитектуры Transformer [8] и большие языковые модели (LLM) стали общедоступными, а многоязычные модели, такие как NLLB-200 [9], поддерживают и кыргызский язык. Однако кыргызский язык по-прежнему относится к языкам с ограниченными цифровыми ресурсами: качество автоматического перевода узкоспециальной технической лексики нестабильно, а языковые модели склонны генерировать правдоподобные, но неверные термины. Поэтому верифицированные отраслевые терминологические базы не теряют значения, а, напротив, становятся ключевым условием надёжного применения ИИ.

Цель статьи — описать проектирование и реализацию электронного словаря технических терминов «Техносөздүк» и обосновать архитектуру его развития с применением технологий искусственного интеллекта.

Анализ программных продуктов для автоматизации перевода. Широкое распространение получили CAT-системы (Computer-Aided Translation — перевод с компьютерной поддержкой), например SDL Trados и Transit, которые существенно облегчают работу переводчиков за счёт памяти переводов и терминологических баз. Отдельный класс образуют системы управления переводами (Translation Management Systems), автоматизирующие переводческие процессы и связанный с ними документооборот [4].

В Кыргызской Республике разработаны и активно применяются следующие основные электронные словари и цифровые платформы для перевода и изучения языков:

– «Эл-Сөздүк» (el-sozduk.kg) — самый популярный и масштабный национальный краудсорсинговый мультиязычный онлайн-словарь и мобильное приложение. База содержит десятки специализированных словарей и более 370 тысяч слов. Включает фундаментальные академические труды (например, кыргызско-русский и русско-кыргызский словари К. К. Юдахина), а также кыргызско-английские направления и разговорники;

– «Э-Сөздүк» от TamgaSoft (tamgasoft.kg) — мультиязычный электронный словарь и онлайн-переводчик. Платформа поддерживает отраслевые базы данных: русско-кыргызский технический словарь, глоссарий банковских и финансовых терминов, словарь компьютерных терминов, а также базу налоговой службы;

– Tili.kg — специализированный электронный ресурс, предлагающий доступ к тематическим словарям. Среди них выделяются Русско-кыргызский словарь компьютерных терминов Microsoft и Русско-кыргызский словарь юридических терминов и иных понятий;

– Кыргызский толковый словарь (разработчик TTDictionary) — популярное мобильное офлайн-приложение для Android и iOS, содержащее более 41 000 статей из академических толковых словарей кыргызского языка. Современные версии приложения также интегрируют ИИ-помощников для генерации контекста и примеров.

Помимо национальных разработок, в госучреждениях, образовательной и корпоративной среде Кыргызстана для работы с международными языками повсеместно применяются глобальные системы электронных словарей, такие как ABBYY Lingvo.

Системы машинного перевода (МП) по функциональному назначению подразделяются на исследовательские прототипы, экспериментальные, промышленные и коммерческие системы. Для решаемой задачи наибольший интерес представляют автоматизированные информационные системы на основе баз данных, относящиеся к системам прямого (словарного) перевода [4].

Среди профессиональных электронных словарей наиболее известен ABBYY Lingvo, отличающийся удобным интерфейсом и многоязычностью. Широко используются также онлайн-сервисы машинного перевода — Google Переводчик (translate.google.com) и Яндекс Переводчик (translate.yandex.ru), которые не требуют установки и поддерживают кыргызский язык.

Таким образом, существующие решения можно условно разделить на три группы: общеязыковые словари, обладающие высоким качеством, но ограниченным покрытием узкой технической лексики; отраслевые глоссарии, не объединённые единой классификацией технических наук; системы машинного перевода общего назначения, обеспечивающие широкое покрытие, но не гарантирующие терминологической точности. Это обосновывает создание специализированного словаря технических терминов с верифицированной базой, который одновременно может служить опорой для ИИ-инструментов.

Описание разработки. Программное обеспечение словаря разработано на языке C# с использованием фреймворка ASP.NET MVC; данные хранятся в СУБД Microsoft SQL Server. Программа зарегистрирована в Кыргызпатенте [7]. Диаграмма вариантов использования системы приведена на рис. 1.



Рис. 1. Диаграмма вариантов использования электронного словаря

Центральным элементом архитектуры ASP.NET MVC является контроллер. Система маршрутизации передаёт входящий запрос соответствующему контроллеру, который обрабатывает данные и возвращает результат. Действия контроллера реализуются публичными методами, сопоставляемыми с запросами.

Контроллер представляет собой класс на языке C#, наследуемый от абстрактного базового класса Controller. В системе реализованы три основных контроллера: AccountController — страница и логика авторизации; HomeController — главная страница и обработка ошибок (методы Index, Privacy и Error); RolesController — управление ролями пользователей. Предусмотрены три роли: администратор, модератор и пользователь.

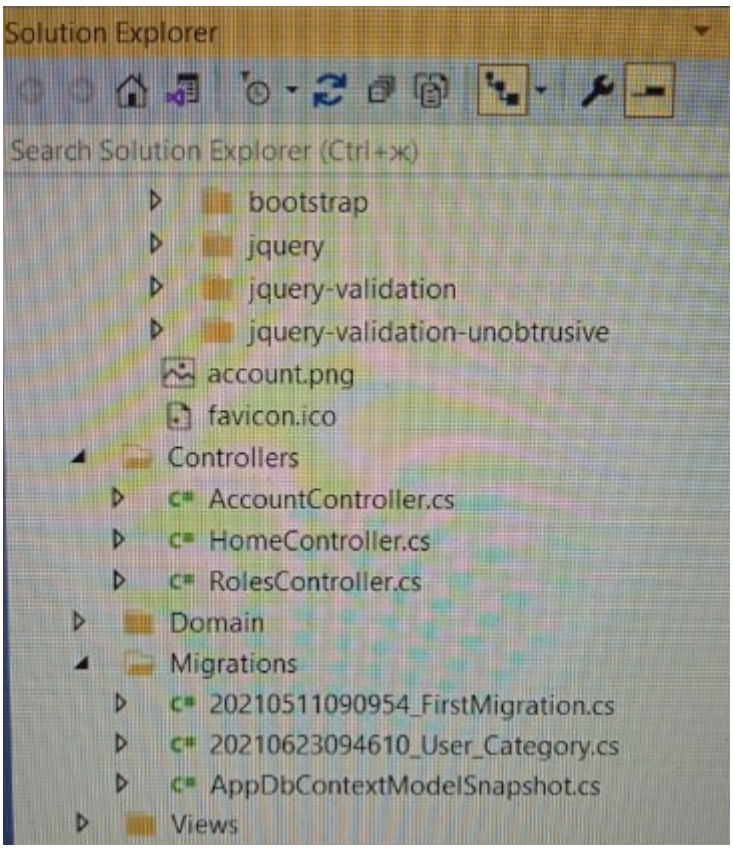
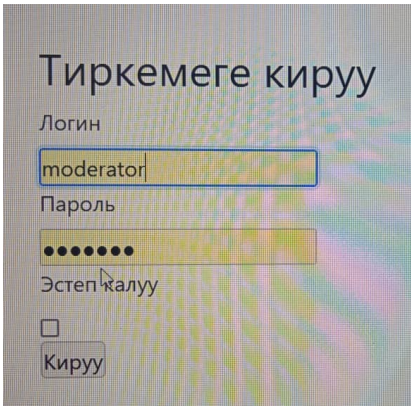


Рис. 2. Окно создания необходимых компонентов системы

Анализ полученных результатов. Разработанная информационная система «Техносөздүк» протестирована и внедрена в КГТУ им. И. Раззакова. Доступ к словарю осуществляется через браузер по адресу sozduk.kstu.kg. Экранные формы системы приведены на рис. 3–6.



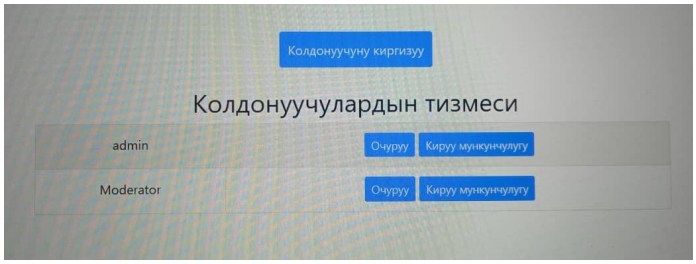


Рис. 3. Окно авторизации пользователей (слева); рис. 4. Окно списка пользователей (справа)

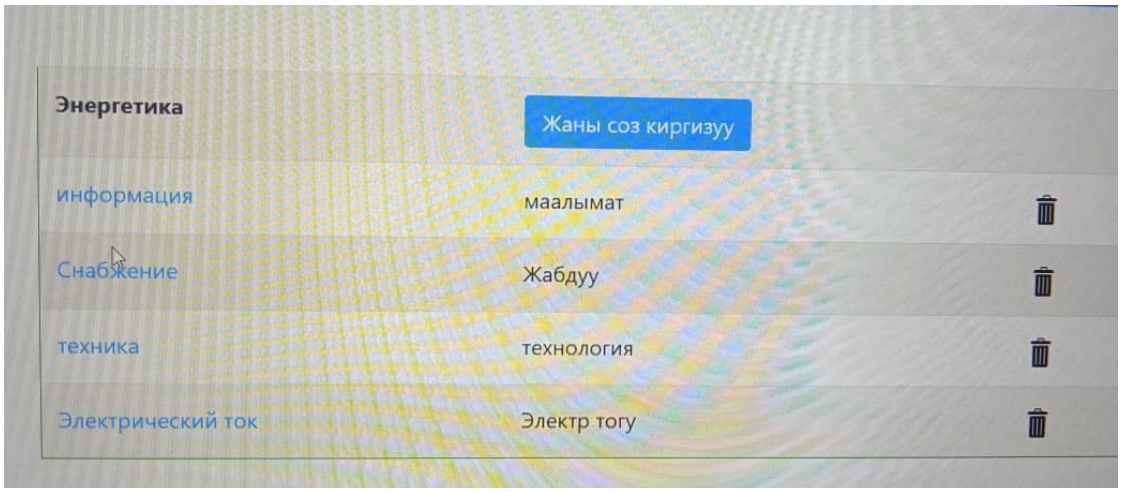


Рис. 5. Окно ввода новых терминов

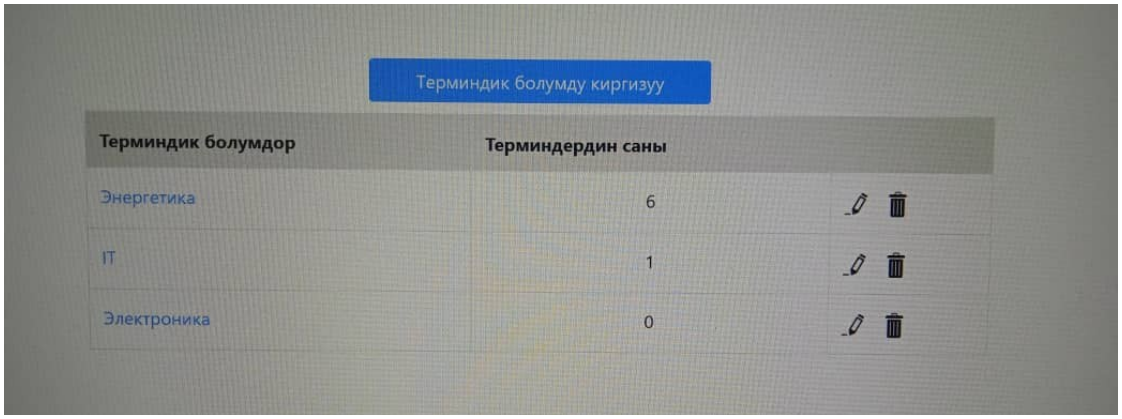


Рис. 6. Окно ввода технических направлений

Перспективы развития словаря на основе технологий ИИ. Опыт эксплуатации показывает, что основным ограничением словаря с поиском по точному совпадению является полнота базы: при отсутствии термина пользователь не получает результата, а пополнение базы требует значительного ручного труда экспертов. Технологии ИИ позволяют снять эти ограничения, не отказываясь от главного преимущества словаря — достоверности. Предлагается дополнить систему модулем интеллектуальной поддержки со следующими функциями.

– Морфологически устойчивый поиск. Лемматизация запросов с учётом аффиксальной морфологии кыргызского языка и нечёткий поиск (с допуском опечаток и вариантов написания букв ө, ү, ң) позволяют находить термин по любой словоформе.

– Семантический поиск. Векторные представления (эмбеддинги) терминов и определений, полученные с помощью многоязычных языковых моделей, обеспечивают поиск по смыслу: пользователь находит термин, даже не зная его точной формы.

– Контекстный перевод по схеме RAG. При переводе предложения или фрагмента текста система извлекает из базы словаря релевантные верифицированные термины и передаёт их языковой модели в качестве обязательного глоссария (Retrieval-Augmented Generation [10]). Это снижает риск терминологических ошибок и «галлюцинаций» модели по сравнению с переводом без опоры на словарь.

– Полуавтоматическое пополнение базы. Для отсутствующих терминов модель формирует кандидаты перевода с пометкой «не верифицировано». Кандидаты, а также термины, автоматически извлечённые из параллельных кыргызско-русских учебных и

нормативных текстов, направляются модератору; в базу вносятся только утверждённые экспертом единицы. Такой порядок естественно опирается на уже реализованную ролевую модель системы (администратор, модератор, пользователь).

– Генерация учебного контента. На основе словарной статьи модель может формировать примеры употребления и тестовые задания, что создаёт основу для планируемой обучающей составляющей словаря.

Принципиальным условием является приоритет верифицированной базы: ответы ИИ должны явно маркироваться как машинные и не смешиваться с экспертно утверждёнными переводами. Технически модуль может быть реализован как отдельный сервис, взаимодействующий с существующим веб-приложением через API, что не требует переработки ядра системы. С учётом требований к защите данных предпочтительно использование открытых моделей, развёрнутых на серверах университета. Для оценки эффективности предлагается сформировать тестовую выборку технических терминов и предложений с эталонным переводом и сравнить режимы «без словаря» и «со словарём» по доле корректно переведённых терминов и автоматическим метрикам качества перевода (BLEU, chrF).

Заключение. Разработан и внедрён в КГТУ им. И. Раззакова двунаправленный кыргызско-русский электронный словарь технических терминов «Техносөздүк», ориентированный на преподавателей, научных сотрудников, студентов, специалистов и всех, кто изучает кыргызский или русский язык как второй. Пользователь может ввести термин или предложение на кыргызском либо русском языке и получить перевод; терминологическая база структурирована по отраслям технических наук, а ролевая модель обеспечивает контроль качества её пополнения. В настоящее время словарь используется отделом государственного языка университета и преподавателями.

Предложенная архитектура модуля ИИ, сочетающая верифицированную терминологическую базу с нейросетевым переводом и большими языковыми моделями по схеме RAG, позволяет перейти от статического словаря к интеллектуальной терминологической платформе: расширить покрытие, обеспечить контекстный перевод и ускорить пополнение базы при сохранении экспертного контроля. Дальнейшие исследования предусматривают реализацию и экспериментальную оценку модуля, подключение обучающей составляющей с интеграцией в автоматизированные системы управления учебным процессом, а также добавление английского и турецкого языков.

Список литературы

1. *Балалаева Е.Ю.* Электронный словарь: сущность, структура, классификации // Современная педагогика. – 2014. – № 4. – URL: https://pedagogika.snauka.ru/2014/04/2238

2. *Бименова Ж.Б., Джумамуратов Р.А., Сидорова Е.А.* Подход к построению русско-казахского тезауруса по информатике // Вестник Бурятского государственного университета. – 2013. – № 9. – С. 53–62.

3. *Ефремова А.Н.* Комплексный электронный словарь: особенности проектирования // Известия Российского государственного педагогического университета им. А.И. Герцена. – 2015. – № 178. – С. 39–43.

4. *Макаревич Т.И.* Автоматизация труда переводчиков и развитие технологий перевода на международном рынке переводов: состояние и тенденции развития // Цифровая трансформация. – 2020. – № 4 (13). – С. 57–67. – https://doi.org/10.38086/2522-9613-2020-4-57-67

5. *Макарич М.В., Попова Ю.Б., Швед М.О.* Лингвистическое и программное обеспечение англо-белорусско-русского словаря технических терминов // Системный анализ и прикладная информатика. – 2018. – № 4. – С. 74–82. – https://doi.org/10.21122/2309-4923-2018-4-74-82

6. *Серова Т.С., Чайникова Г.Р.* Содержание поэтапной разработки учебного терминологического словаря тезаурусного типа // Педагогическое образование в России. – 2015. – № 2. – С. 143–151.

7. *Торобеков Б.Т., Джунусалиев Н.Дж., Тойбаева Ж.Дж., Асанбекова Ч.М.* Web-приложение «Электронный словарь технических слов «Techno сөздүк»: авторское свидетельство на программу для ЭВМ № 757. – Бишкек: Кыргызпатент, 03.01.2022.

8. *Vaswani A., Shazeer N., Parmar N. et al.* Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.

9. *NLLB Team, Costa-jussà M.R., Cross J. et al.* No Language Left Behind: Scaling Human-Centered Machine Translation. – arXiv:2207.04672. – 2022.

10. *Lewis P., Perez E., Piktus A. et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459–9474.

Machine-Learning-Driven Adaptive Traffic Management for Multi-Cluster Kubernetes Service Mesh via Gateway API

Nurdaulet Kemelbekov*, Diana R. Rakhimova*
*Al-Farabi Kazakh National University
Almaty, Kazakhstan
Email: kemellbekov@gmail.com

***Abstract*—Modern banking infrastructures increasingly rely on multi-datacenter Kubernetes platforms in which service mesh technologies such as Istio and the Kubernetes Gateway API govern inter-service traffic. Existing traffic-shifting mechanisms are largely static or threshold-based, reacting slowly to gradual quality degradation and requiring manual tuning by platform engineers. This paper proposes an ML-controller architecture that continuously fuses Envoy/Prometheus metrics, application log signals from OpenSearch, and Kubernetes pod-state data into a per-backend health score, which is then used to adjust `HTTPRoute` traffic weights automatically. We describe the system architecture, the experimental test stand built on Kubernetes, Istio, Prometheus, Grafana, Jaeger and OpenTelemetry, and an evaluation methodology comparing the proposed approach against static weighted routing and threshold-based canary analysis.**

***Index Terms*—Kubernetes, service mesh, Gateway API, Istio, traffic management, machine learning, DevOps, canary deployment, observability**

I. Introduction

Container orchestration platforms based on Kubernetes have become the operational backbone of large-scale, regulated environments such as banking IT infrastructure, where workloads are typically distributed across several datacenters for fault tolerance and disaster recovery. In such multi-cluster deployments, a service mesh layer (e.g., Istio, built on the Envoy proxy) is commonly used to manage east-west and north-south traffic, enforce mutual TLS, collect telemetry, and implement progressive delivery patterns such as canary releases and blue-green deployments.

Traffic distribution between service versions or between datacenters is conventionally controlled through the Kubernetes Gateway API's `HTTPRoute` resource, which allows an operator to assign fixed relative weights to backend references. In practice, these weights are either set manually by DevOps/SRE engineers or adjusted by threshold-based progressive-delivery controllers that evaluate a small set of fixed thresholds (e.g., error rate $> 1\%$ over a fixed window). Such approaches have two structural limitations. First, they react only after a threshold is breached, which is often too late to prevent user-facing impact in latency-sensitive financial transactions. Second, they typically consider a narrow slice of the available telemetry (usually only metrics), ignoring log-level error signatures and Kubernetes-level pod health signals that frequently precede metric-level degradation.

This paper presents the design of an ML-controller that treats traffic weighting as a continuous, learned control problem rather than a static configuration. The controller ingests multiple telemetry sources, produces a predictive per-backend health score, and translates that score into safe, rate-limited adjustments of Gateway API routing weights. The contributions of this work are threefold:

- an architecture for multi-source telemetry fusion (metrics, logs, cluster state) for service mesh traffic control;
- a closed-loop control scheme with explicit safety bounds, designed for regulated banking environments;
- an experimental test stand and evaluation methodology for comparing learned traffic management against static and threshold-based baselines.

II. Related Work

Service mesh traffic management has traditionally relied on declarative routing rules. The Kubernetes Gateway API standardizes the `HTTPRoute` object, in which weighted traffic splitting across backend services is expressed through weights on backend references [1], as a standardized alternative to the earlier, mesh-specific Istio `VirtualService`/`DestinationRule` model [2]. Progressive-delivery controllers such as Flagger [3] and Argo Rollouts [4] automate canary promotion by polling a small number of metrics (typically request success rate and latency percentiles) against fixed thresholds during a fixed analysis window, then incrementally shifting the weight if the thresholds are satisfied. The Service Mesh Interface specification [5] attempted to standardize this class of traffic-split abstractions across mesh implementations. The architectural principles of the service mesh, its performance overhead, and its practical use in Kubernetes were examined in our earlier work [6].

Separately, the networking research community has explored learning-based control for adjacent problems.

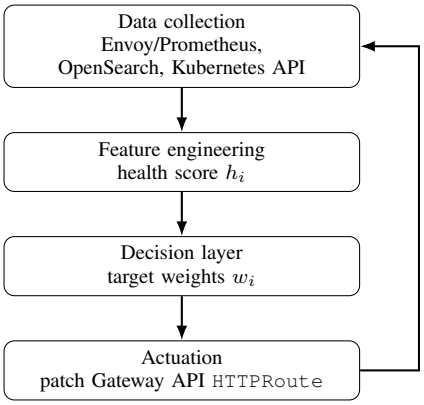


Fig. 1. Closed-loop architecture of the ML-controller.

Knowledge-defined networking [7] proposed closing the network control loop with a knowledge plane built on machine learning and analytics. Reinforcement-learning approaches have been applied to related control problems such as congestion control, where a learned policy can outperform hand-tuned heuristics under diverse and shifting network conditions [8]. These works motivate treating service-level traffic weighting, like congestion control, as a control problem that benefits from continuous, data-driven adaptation rather than static thresholds, but to our knowledge they have not been applied directly to Gateway API-based Kubernetes traffic splitting with multi-source telemetry fusion, which is the gap this work addresses.

III. Proposed Architecture

The proposed ML-controller is organized into four layers, illustrated conceptually in Fig. 1.

A. Data Collection Layer

Three independent telemetry sources are polled on a fixed interval:

- **Metrics**: Envoy sidecar and Prometheus time series, including per-backend request rate, P50/P95/P99 latency, and 4xx/5xx error ratios;
- **Logs**: application and access logs indexed in OpenSearch, from which error-signature frequency and novel error-type occurrence are extracted;
- **Cluster state**: Kubernetes API data on pod readiness, restart counts, and resource pressure (CPU/memory throttling) for each backend's pods.

B. Feature Engineering and Health Scoring

Raw signals are aggregated per backend over a sliding window and normalized into a feature vector $\mathbf{x}_i$ for backend i. A lightweight model f_θ maps $\mathbf{x}_i$ to a scalar predicted health score $h_i \in [0, 1]$, as given by (1):

$$\begin{aligned} h_i &= f_\theta(\mathbf{x}_i), \\ \mathbf{x}_i &= \left[\, \text{latency}_i,\ \text{err\_rate}_i,\ \text{log\_sig}_i,\ \text{pod\_state}_i \,\right] \end{aligned} \tag{1}$$

The model is intentionally kept simple (e.g., gradient-boosted trees or a small online-learned regressor) to keep inference latency low enough for the control loop and to keep the decision path auditable, which is important for change-management processes in a banking environment.

C. Decision and Actuation Layers

Predicted health scores across all candidate backends are converted into target traffic weights w_i using a softmax-style normalization bounded by configurable minimum and maximum weight constraints $w_{\min}$ and $w_{\max}$, as given by (2):

$$w_i = \text{clip}\left(\frac{\exp(\beta h_i)}{\sum_j \exp(\beta h_j)},\ w_{\min},\ w_{\max} \right) \tag{2}$$

where β controls how aggressively weight shifts respond to health-score differences. To avoid oscillation, weight changes between consecutive control cycles are rate-limited, and a rollback rule reverts traffic fully to the last known-good backend if a hard safety threshold (e.g., error rate spike) is breached regardless of the model's output. The resulting weights are applied by patching the Gateway API `HTTPRoute` object in the `gateway-app` namespace, with the controller itself running in a dedicated `ml-system` namespace.

D. Closed-Loop Operation

The four layers execute cyclically, forming a closed control loop: observe (collect telemetry) → predict (health score) → decide (target weights) → act (patch `HTTPRoute`) → observe again. This design keeps the ML component isolated from the data plane: Envoy continues to route traffic according to the last applied `HTTPRoute` configuration even if the controller itself becomes unavailable, which bounds the blast radius of a controller failure.

IV. Experimental Test Stand

To evaluate the architecture, a dedicated test stand was built, independent of production banking infrastructure. It consists of a Kubernetes cluster running:

- **Istio** as the service mesh, exposing Gateway API-compatible routing;
- **Prometheus** [9] and **Grafana** for metrics collection and visualization;
- **Jaeger** and **OpenTelemetry** [10] for distributed tracing;
- an **OpenSearch**-based logging pipeline for application log aggregation.

The test stand is a single Kubernetes cluster; multi-cluster and datacenter-to-disaster-recovery scenarios are outside the scope of the present evaluation (see Section VII). The ML-controller itself is implemented as a research prototype with a minimal dependency footprint, reflecting its role as an experimental component rather than a hardened production service at this stage.

TABLE I
COMPARED ROUTING STRATEGIES

	Static	Threshold-based	ML-controller
Feedback	None	Yes, within a fixed analysis window	Yes, continuous (closed loop)
Input signals	None	Request success rate, latency percentiles	Metrics, logs, cluster state
Decision rule	Fixed weights	Fixed thresholds	Health score h_i, softmax-style weights
Weight update	None	Incremental promotion or rollback	Rate-limited, bounded by $w_{\min}$, $w_{\max}$

V. EVALUATION METHODOLOGY

We propose to compare three routing strategies under synthetic fault injection (elevated latency, partial error injection, and pod resource throttling applied to a subset of backend replicas):

1) **Static routing** — fixed weighted round-robin, no feedback;
2) **Threshold-based canary analysis** — Flagger-style fixed-threshold promotion/rollback;
3) **ML-controller** — the proposed continuous health-score-driven weighting.

Table I summarizes the main characteristics of the three strategies. The comparison metrics are: (i) time-to-detect a degraded backend, (ii) P95/P99 end-to-end latency during the fault window, (iii) proportion of requests routed to the degraded backend during the fault window, and (iv) stabilization time after the fault is resolved. Quantitative results for these metrics will be obtained by executing the fault-injection experiments on the test stand described in Section IV; they are not reported in this paper.

VI. DISCUSSION

From a DevOps/SRE perspective, the main expected value of the proposed approach is reducing the manual toil of tuning canary thresholds per service and shortening the detection-to-mitigation interval for gradually developing incidents that do not immediately breach a fixed threshold. Because the controller only ever adjusts weights within configured bounds and can be disabled without affecting the last-applied routing state, it is designed to be introduced incrementally into an existing GitOps-managed platform: it can run in a shadow (recommend-only) mode initially, with a human operator approving weight changes via the existing CI/CD pipeline before being granted direct write access to `HTTPRoute` objects. This staged rollout is particularly relevant for banking environments, where change-management and auditability requirements constrain how much autonomous control can be delegated to an automated system, and where model decisions need to remain explainable to satisfy internal risk and compliance review.

VII. CONCLUSION AND FUTURE WORK

This paper presented the architecture of an ML-controller for adaptive, multi-source-telemetry-driven traffic management in a multi-cluster Kubernetes service mesh, targeting the Gateway API as its actuation interface. The design fuses metrics, logs, and cluster state into a per-backend health score used to continuously and safely adjust routing weights, in contrast to static or threshold-based canary approaches. Future work includes completing the fault-injection evaluation on the described test stand, comparing several candidate model families for the health-scoring function, extending the approach to multi-datacenter failover scenarios between a primary datacenter (DC) and a disaster-recovery (DR) site, and studying the explainability of the controller's decisions for compliance review in a regulated banking context.

ACKNOWLEDGMENT

This work was conducted as part of doctoral studies in program 8D06101 at Al-Farabi Kazakh National University.

REFERENCES

[1] Kubernetes SIG-Network, "HTTP traffic splitting," Gateway API documentation. Accessed: Oct. 1, 2026. [Online]. Available: https://gateway-api.sigs.k8s.io/guides/traffic-splitting/
[2] Istio Authors, "Traffic management," Istio documentation. Accessed: Oct. 1, 2026. [Online]. Available: https://istio.io/latest/docs/concepts/traffic-management/
[3] Flux Project (CNCF), "Flagger documentation." Accessed: Oct. 1, 2026. [Online]. Available: https://docs.flagger.app/
[4] Argo Project, "Argo Rollouts documentation." Accessed: Oct. 1, 2026. [Online]. Available: https://argoproj.github.io/argo-rollouts/
[5] Service Mesh Interface Authors, "SMI: A standard interface for service meshes on Kubernetes," CNCF archived project. Accessed: Oct. 1, 2026. [Online]. Available: https://smi-spec.io/
[6] N. K. Kemelbekov, D. R. Rakhimova, and E. Adali, "Service Mesh: Architectural principles, performance and practical use in microservice systems" (in Russian), *Herald of the Kazakh-British Technical University*, vol. 23, no. 2, pp. 276–289, 2026, doi: 10.55452/1998-6688-2026-23-2-276-289.
[7] A. Mestres *et al.*, "Knowledge-defined networking," *ACM SIGCOMM Computer Communication Review*, vol. 47, no. 3, pp. 2–10, 2017.
[8] N. Jay, N. H. Rotman, P. B. Godfrey, M. Schapira, and A. Tamar, "A deep reinforcement learning perspective on Internet congestion control," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, 2019, vol. 97, pp. 3050–3059.
[9] Cloud Native Computing Foundation, "Prometheus documentation." Accessed: Oct. 1, 2026. [Online]. Available: https://prometheus.io/docs/
[10] Cloud Native Computing Foundation, "OpenTelemetry documentation." Accessed: Oct. 1, 2026. [Online]. Available: https://opentelemetry.io/docs/

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ОРГАНИЗАЦИИ НАУЧНОГО ПОИСКА: ПОДХОДЫ К ИССЛЕДОВАНИЮ НЕЙРОМОРФНЫХ СВОЙСТВ И ВЫЧИСЛИТЕЛЬНОГО ПОТЕНЦИАЛА ПОЛИМЕРОВ

ARTIFICIAL INTELLIGENCE IN ORGANIZING SCIENTIFIC INQUIRY: APPROACHES TO STUDYING THE NEUROMORPHIC PROPERTIES AND COMPUTATIONAL POTENTIAL OF POLYMERS

Кабдушев Ш. Б., Сергеев А. С., Нуралиев Ф. М., Байпакбаева С. Т., Мун Г. А.

Аннотация. Рассматривается изменение организации научного поиска по мере объединения выдвижения гипотез, моделирования, отбора кандидатов и экспериментальной проверки в единый исследовательский цикл с участием искусственного интеллекта. На опубликованных примерах показано, какие части такого цикла уже реализованы в разных областях. Мы рассматриваем, как этот подход можно применить к изучению нейроморфных свойств и вычислительного потенциала полимеров. Опираясь на опубликованные исследования, мы обосновываем подход, при котором предполагаемая функция материала определяет выбор модели и способ её проверки. Преимущества ИИ рассматриваются вместе с рисками ошибочной интерпретации и утраты независимости проверки. Показана взаимная связь между применением ИИ в исследовании материалов и поиском более экономичной вычислительной основы самого ИИ. Такая организация исследований подготавливает научную среду к применению более самостоятельных интеллектуальных систем. Это относится и к общему искусственному интеллекту (AGI), если он будет создан.

Ключевые слова: искусственный интеллект, организация научного поиска, моделирование, нейроморфные полимеры, вычислительный потенциал, энергоэффективность, AGI.

Abstract. This conceptual review examines changes in scientific inquiry as hypothesis generation, modelling, candidate selection and experimental validation are combined into a single research cycle involving artificial intelligence. Published examples show which parts of this cycle are already implemented across different fields. We consider how this approach can be applied to studying the neuromorphic properties and computational potential of polymers. Drawing on published research, we propose an approach in which the intended function of a material guides model selection and the design of validation. We discuss the benefits of AI alongside the risks of misinterpretation and loss of independent validation. We also examine the reciprocal relationship between using AI in materials research and seeking a more economical computing basis for AI itself. Such research practices can prepare science for more autonomous intelligent systems, including artificial general intelligence (AGI), should it be developed.

Keywords: artificial intelligence, scientific inquiry, modelling, neuromorphic polymers, computational potential, energy efficiency, AGI.

Введение

Развитие искусственного интеллекта уже меняет сам способ организации научного поиска. Анализ литературы, выдвижение гипотез, моделирование, выбор кандидатов и эксперимент всё чаще связываются так, что результат одного этапа определяет следующий шаг. Такие методы пока реализованы фрагментарно и различаются по степени самостоятельности программ. Тем не менее они позволяют обсуждать изменение исследовательской работы уже на основе практических примеров. Для науки важно понять, при каких условиях эта связь помогает получать обоснованное знание и где возникают новые источники ошибок.

Мы рассматриваем эти изменения на примере планируемого исследования нейроморфных свойств и вычислительного потенциала полимеров. Наш интерес связан с возможностью использовать поведение материала для обработки информации. Здесь необходимо пройти путь от физических свойств и предполагаемой функции к математическому описанию, расчётному прогнозу и его проверке на реальном образце. На примере исследования полимеров можно показать, как ИИ помогает связывать физическое описание материала, моделирование и экспериментальную проверку.

Цель статьи – обосновать логику такого подхода, опираясь на опубликованные исследования, и определить его возможности и ограничения. Мы объясняем, почему выбрали полимеры и как планируем сочетать их моделирование с экспериментальной проверкой, а затем рассматриваем значение этого подхода для научной среды. Далее мы обсуждаем, как эта организация исследований могла бы развиваться при появлении общего искусственного интеллекта (AGI).

1. От отдельных инструментов к единому исследовательскому циклу

Пересмотр исходной гипотезы с учётом полученных результатов – обычная часть научного исследования. Развитие ИИ расширяет возможности программ участвовать в выборе следующего шага. Существенным становится переход от отдельного расчёта к последовательности решений: что предположить, какую модель построить, что проверить и как изменить исходное объяснение. Публикации из разных областей позволяют увидеть отдельные реализации этой логики.

Один из этапов такого поиска – выбор объектов для дальнейшего исследования. Здесь ИИ помогает определить, какие из них стоит проверять в первую очередь. Например, AlphaFold 3 прогнозирует пространственную структуру комплексов белков и других биомолекул [1]. Такой прогноз помогает выделить взаимодействия, заслуживающие дальнейшей проверки. В материаловедении GNoME использует машинное обучение для поиска кристаллических структур, перспективных по расчётным критериям устойчивости [2]. В обоих случаях программа сужает область последующего поиска. Вычислительный прогноз требует дальнейшей

проверки: найденная структура ещё не равнозначна изготовленному и исследованному материалу.

Более тесная связь с физическим экспериментом реализована в автономной лаборатории A-Lab. Установка выбирает условия синтеза, выполняет операции, анализирует продукты и использует результаты для следующих попыток. Согласно исправленной публикации, за 17 суток были получены 36 соединений из 57 целевых [3, 4]. Этот пример показывает, как программа использует результаты эксперимента для выбора условий следующей попытки.

Объединять несколько этапов исследования помогают и научные агентные системы – программы, которые последовательно используют разные инструменты. Co-Scientist предлагает и уточняет гипотезы во взаимодействии с учёными; часть предложений проверена в лаборатории [5]. The AI Scientist объединяет разработку идей, программирование, вычислительные эксперименты и подготовку текста в исследованиях по машинному обучению [6]. Эти работы относятся к разным задачам и не образуют единую универсальную систему. Вместе они показывают, что участие ИИ уже может охватывать несколько взаимосвязанных исследовательских операций.

Из этих примеров мы выделяем принцип, существенный для полимерного исследования: каждый расчёт или опыт должен давать основание для уточнения дальнейшего поиска. Его применение требует конкретного вопроса, измеримых характеристик и способа проверки. В нашем случае это приводит к вопросу о том, какие свойства полимеров могут обеспечивать обработку информации и почему поиск таких свойств важен для развития ИИ.

2. Почему мы исследуем вычислительный потенциал полимеров

Выбор этой задачи связан в том числе с потребностями самого ИИ. Его развитие зависит от доступности вычислений и энергии; с ростом нагрузки на вычислительную инфраструктуру возрастает потребность в энергоэффективном оборудовании [7]. Один из исследуемых путей – выполнять часть операций непосредственно там, где хранится информация, сокращая её передачу между памятью и вычислительными узлами. Один из примеров реализации этого принципа – аналоговый чип для распознавания речи [8]. Эта работа показывает, как можно изменить физическую организацию вычислений, хотя сама относится к другой, неполимерной технологии.

Среди возможных вычислительных материалов рассматриваются полимеры – вещества, состоящие из длинных молекулярных цепей. В опубликованных исследованиях взаимодействия в водных растворах некоторых полимеров рассматриваются с точки зрения образования аналогов нейронных сетей [9, 10]. В основе этой идеи лежит зависимость состояния отдельных элементов системы от внешнего воздействия и взаимного влияния элементов. Термин «нейроморфный» обозначает здесь

сходство отдельных принципов поведения материала с работой нервных систем, например сохранение влияния предшествующих воздействий или изменение связей.

Опубликованные работы включают обзор научных идей, теоретическую модель и экспериментальные исследования [9-11]. Обзор 2025 года развивает представление о полимерных системах как носителях информационных процессов [9]. Работа 2026 года предлагает теоретическое описание коллективного поведения раствора полимера с кислотными группами через взаимодействия электрических зарядов [10]. В экспериментальном исследовании полимерных растворов на основе измерений предложено объяснение того, как участки полимерных цепей объединяются в более крупные структуры [11]. Физические измерения и нейросетевая интерпретация выполняют здесь разные задачи: первые описывают материал, вторая задаёт направление исследования его возможных функций.

Опираясь на эти работы, мы ставим следующий вопрос: при каких условиях описанное поведение способно обеспечить воспроизводимое выполнение вычислительной задачи? Способность материала реагировать на сигнал или сохранять след воздействия ещё нужно связать с тем, какую информацию мы вводим, какой результат получаем и как оцениваем его правильность. Мы предлагаем выяснить, какие физические свойства материала необходимы для выполнения выбранной вычислительной задачи.

Практическая перспектива состоит в том, что подходящая полимерная система могла бы стать основой более экономичных вычислений для отдельных задач ИИ. Однако потенциальные преимущества зависят от точности и долговечности системы, а также от затрат на её изготовление, управление и считывание результатов. Поэтому снижение стоимости и энергопотребления мы рассматриваем как цель дальнейшего исследования всей системы. ИИ может помогать исследовать полимеры, а новые полимерные материалы потенциально могут расширить возможности самого ИИ.

3. Как мы планируем связать моделирование и проверку

Первый шаг исследования – определить вычислительную задачу, которую должна выполнять система. Например, можно поставить задачу различения последовательностей сигналов или восстановления информации по неполным входным данным. После этого требуется определить, какие изменения состояния материала должны обеспечивать нужный результат и какие величины можно измерить. Такая последовательность позволяет сопоставлять расчётную модель с заранее сформулированным требованием к её поведению.

Мы планируем использовать ИИ, чтобы связать анализ литературы, построение модели и выбор экспериментальной проверки. При работе с литературой он может помогать сопоставлять физические объяснения и выявлять различия в исходных допущениях. При построении модели

материала – готовить расчётные инструменты и сравнивать последствия выбранных предположений. При отборе вариантов – выбирать условия, при которых разные гипотезы предсказывают разные результаты, чтобы эксперимент помог выбрать наиболее обоснованное объяснение. При этом модель ИИ, предлагающая решение, и физическая модель полимера остаются разными инструментами, каждый из которых требует собственной проверки.

Мы планируем сравнивать варианты материалов, условия воздействия и способы описания поведения системы. Сначала оценивается, насколько каждый вариант соответствует выбранной функции, затем формулируется проверяемый прогноз. Для экспериментальной проверки необходимо выбрать измерение, результаты которого можно сопоставить с прогнозом. Результаты измерений следует использовать для уточнения модели и выбора следующего опыта. Неудачный опыт в таком случае помогает уточнить границы применимости объяснения и отказаться от необоснованного направления поиска.

Для нас принципиально сохранять возможность пересмотра на каждом этапе. Если модель выполняет вычислительную задачу лишь при допущении, которое не удаётся связать с измеримым свойством материала, нужно уточнять и модель, и постановку задачи. Если разные объяснения одинаково согласуются с известными данными, требуется дополнительная проверка, позволяющая их различить. Обратная связь позволяет пересматривать исходные предположения и уточнять дальнейший план исследования.



Рисунок 1 – Предлагаемая связь моделирования и экспериментальной проверки в исследовании вычислительного потенциала полимеров. Схема отражает планируемую организацию работы.

Для решения этой задачи необходимо совместно использовать знания о химических свойствах полимеров, методы физического моделирования и теорию вычислений. ИИ может помогать специалистам сопоставлять данные и методы этих областей. Пользу ИИ следует оценивать по тому, помогает ли он обосновать выбор вариантов и предложить способ их

проверки. Это позволяет перейти от общих ожиданий эффективности к конкретным преимуществам и рискам предлагаемого подхода.

4. Возможности ИИ и условия достоверности исследования

Главное преимущество единого исследовательского цикла мы видим в более обоснованном распределении исследовательских усилий. Сопоставление литературы и расчётных вариантов помогает выделять кандидатов для проверки, а результаты опытов – уточнять критерии отбора. Для полимерной задачи это означает возможность сосредоточиться на свойствах, существенных для выбранной функции. ИИ также может помогать проверять, согласуются ли предположения специалистов одной области с ограничениями, известными в другой.

Однако ошибка на одном этапе может повлиять на все последующие решения. Неверно прочитанная работа может стать основанием физической модели, а ошибка в расчётной программе – повлиять на выбор образца. Убедительное объяснение, предложенное языковой моделью, способно создать ощущение понимания без достаточного обоснования [12]. Поэтому независимость проверки должна быть заложена в подход: повторный ответ той же системы сам по себе не служит независимым подтверждением.

Таблица 1 – Участие ИИ на этапах полимерного исследования и условия проверки

Этап исследования	Ожидаемая помощь ИИ	Условие проверки
Литература и гипотезы	Сопоставить объяснения и возможные связи	Проверить первоисточники и границы их выводов
Модель и отбор вариантов	Подготовить расчёты и сравнить прогнозы	Проверить допущения и программу на случаях с известным решением
Эксперимент и уточнение	Сопоставить измерения с прогнозом	Сохранить исходные данные. Повторить измерения, на которых основаны основные выводы

При планировании исследования необходимо предусмотреть проверку альтернативных гипотез и учитывать результаты, которые не подтверждают исходную идею. Если отбор опирается преимущественно на привычные объекты и успешные примеры, поиск рискует замкнуться на уже известных направлениях. Анализ научных публикаций выявил связь применения ИИ с ростом влияния отдельных исследователей и одновременно с сужением круга исследуемых тем [13]. Эта наблюдаемая связь не доказывает неизбежность такого эффекта, но даёт основания уделять внимание альтернативным гипотезам и учитывать отрицательные результаты.

Чем самостоятельнее действует программа, тем важнее сохранять данные, позволяющие восстановить ход исследования и основания

принятых решений. Для этого нужно сохранять версии моделей, условия расчётов и исходные измерения. При подключении оборудования необходимы ограничения допустимых действий и возможность остановки; при работе с неопубликованными данными – защита информации и согласованные условия доступа [14]. Эти требования определяют, какие операции можно поручить системе и где требуется дополнительная проверка. Их выполнение важно и для одного проекта, и для научной среды, использующей всё более самостоятельные инструменты.

5. Исследовательская среда и перспектива AGI

Чтобы проверить гипотезу с помощью моделирования и эксперимента, нужны доступные данные, согласованные способы их описания, средства управления экспериментом и понятные критерии проверки. В исследовании полимеров это означает необходимость согласовать данные о материале, расчётную модель и методику измерений. Это облегчает взаимодействие исследователей разных специальностей и делает результаты отдельных этапов пригодными для дальнейшего использования.

Мы рассматриваем распространение таких процессов как существенную часть изменений, которые ИИ вносит в науку. По мере их развития более самостоятельная программа получает возможность включаться в уже организованную последовательность исследовательских действий. Если появится AGI, способный решать широкий круг задач и переносить знания между областями, его применение также сможет опираться на эту среду. В этом смысле применение AGI могло бы продолжить развитие уже существующих способов организации исследований.

Из этого не следует, что AGI обязательно появится или что его внедрение окажется простым. При этом уже для современных систем важны качество данных, проверяемость моделей, сохранение альтернатив и ответственность за цели исследования. Для нас обсуждение AGI связано прежде всего с тем, как организовать исследования с учётом будущего развития систем ИИ. Организация этой работы сегодня будет влиять на то, какую пользу эти системы смогут принести исследованиям в будущем.

Заключение

Рассмотренные примеры показывают, как ИИ включается в последовательности исследовательских решений, связывая гипотезы, моделирование, отбор и проверку. Для исследования нейроморфных свойств полимеров мы предложили подход, в котором предполагаемая вычислительная функция определяет требования к физической модели и будущему эксперименту. Опубликованные работы о полимерных системах служат научной основой для постановки этой задачи; её дальнейшее развитие требует конкретизации моделей и критериев проверки.

Ценность предлагаемой организации поиска состоит в возможности использовать результаты каждого этапа для уточнения модели и выбора дальнейших действий. Преимущества ИИ реализуются, если результаты проверяются независимо, а основания исследовательских решений можно восстановить. Исследование полимеров позволяет рассмотреть взаимное влияние науки и ИИ: новые методы помогают исследовать материалы, которые потенциально способны изменить вычислительную основу этих методов. Распространение подобных подходов формирует научную среду, в которой применение AGI, если он появится, может стать продолжением уже начавшихся изменений в организации исследований.

Литература

1. Abramson J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 // Nature. 2024. Vol. 630. P. 493–500. DOI: 10.1038/s41586-024-07487-w.

2. Merchant A. et al. Scaling deep learning for materials discovery // Nature. 2023. Vol. 624. P. 80–85. DOI: 10.1038/s41586-023-06735-9.

3. Szymanski N. J. et al. An autonomous laboratory for the accelerated synthesis of inorganic materials // Nature. 2023. Vol. 624. P. 86-91. DOI: 10.1038/s41586-023-06734-w. С учётом исправления [4].

4. Szymanski N. J. et al. Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials // Nature. 2026. DOI: 10.1038/s41586-025-09992-y.

5. Gottweis J. et al. Accelerating scientific discovery with Co-Scientist // Nature. 2026. Vol. 655. P. 487-496. DOI: 10.1038/s41586-026-10644-y.

6. Lu C. et al. Towards end-to-end automation of AI research // Nature. 2026. Vol. 651. P. 914-919. DOI: 10.1038/s41586-026-10265-5.

7. International Energy Agency. Energy and AI. Paris: IEA, 2025. URL: https://www.iea.org/reports/energy-and-ai (дата обращения: 02.10.2026).

8. Ambrogio S. et al. An analog-AI chip for energy-efficient speech recognition and transcription // Nature. 2023. Vol. 620. P. 768-775. DOI: 10.1038/s41586-023-06337-5.

9. Kabdushev S., Gabrielyan O., Kopishev E., Suleimenov I. Neural network properties of hydrophilic polymers as a key for development of the general theory of evolution // Royal Society Open Science. 2025. Vol. 12. Art. 242149. DOI: 10.1098/rsos.242149.

10. Kabdushev S., Shaltykova D., Kopishev E., Seitenova G., Dyusssova R., Suleimenov I. Polyacid Solutions as an Analogue of a Neural Network // Polymers. 2026. Vol. 18. Art. 279. DOI: 10.3390/polym18020279.

11. Kabdushev S., Mun G., Suleimenov I., Alikulov A., Shaikhutdinov R., Kopishev E. Formation of Hydrophobic–Hydrophilic Associates in the N-Vinylpyrrolidone and Vinyl Propyl Ether Copolymer Aqueous Solutions // Polymers. 2023. Vol. 15. Art. 3578. DOI: 10.3390/polym15173578.

12. Messeri L., Crockett M. J. Artificial intelligence and illusions of understanding in scientific research // Nature. 2024. Vol. 627. P. 49–58. DOI: 10.1038/s41586-024-07146-0.

13. Hao Q., Xu F., Li Y., Evans J. Artificial intelligence tools expand scientists' impact but contract science's focus // Nature. 2026. Vol. 649. P. 1237–1243. DOI: 10.1038/s41586-025-09922-y.

14. Tang X. et al. Risks of AI scientists: prioritizing safeguarding over autonomy // Nature Communications. 2025. DOI: 10.1038/s41467-025-63913-1.

КОНТЕЙНЕРИЗАЦИЯ И ОРКЕСТРАЦИЯ КАК ОСНОВА CLOUD-NATIVE РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Жамбылхан Санжар Еркінұлы
Таразский региональный университет имени М. Х. Дулати, Республика Казахстан, г. Тараз
Почта: 707bma@gmail.com

Аннотация.
В статье рассматриваются контейнеризация и оркестрация как технологическая основа cloud-native разработки программного обеспечения. Проведено сопоставление виртуальных машин и контейнеров, описана связь контейнеров с микросервисной архитектурой, рассмотрены принципы работы платформы Kubernetes и роль контейнеров в процессах DevOps и CI/CD. Выявлены основные риски безопасности и ограничения технологии. На примере архитектуры интернет-магазина показано, как независимое масштабирование сервисов и автоматизация развёртывания повышают устойчивость системы. Сформулированы рекомендации по внедрению контейнерной платформы.

Введение

Современные программные системы всё реже представляют собой единое приложение, работающее на одном сервере. Веб-сервисы, серверные части мобильных приложений, платформы электронной коммерции и аналитические комплексы состоят из множества взаимодействующих компонентов, которые необходимо часто обновлять, масштабировать и восстанавливать после сбоев. Облачные платформы, определяемые как модель сетевого доступа по требованию к общему пулу настраиваемых вычислительных ресурсов [1], решили задачу быстрого получения инфраструктуры, однако не отвечают на вопрос, как единообразно упаковывать, доставлять и запускать само программное обеспечение. Ответом на этот вопрос стали контейнеры и системы оркестрации.

Массовое распространение контейнерной технологии связано с появлением в 2013 году платформы Docker, которая упростила создание, распространение и запуск изолированных приложений и позволила использовать одинаковую среду на компьютере разработчика, в тестовом контуре и в промышленной эксплуатации [2]. Следующим шагом стало появление систем управления большим числом контейнеров, среди которых наибольшее распространение получила платформа Kubernetes.

Актуальность темы определяется тем, что контейнеризация и оркестрация превратились в стандартную основу разработки и эксплуатации распределённых систем. Знание принципов их работы необходимо не только инженерам по эксплуатации, но и разработчикам, поскольку архитектура приложения напрямую зависит от среды, в которой оно будет выполняться.

Цель статьи – рассмотреть контейнеризацию и оркестрацию как технологическую основу cloud-native разработки программного обеспечения, определить их преимущества и ограничения и показать практическое применение на примере типовой архитектуры веб-приложения. Для достижения цели решаются следующие задачи: проанализировать различия между виртуальными машинами и контейнерами; рассмотреть связь контейнеров с микросервисной архитектурой; описать основные принципы работы Kubernetes; оценить роль контейнеров в процессах DevOps и CI/CD; выявить основные риски безопасности; сформулировать рекомендации по внедрению. Исследование выполнено на основе анализа научных публикаций, публикаций стандартизирующих организаций и технической документации.

Проблема воспроизводимости среды и переход к контейнерам

Классическая проблема разработки программного обеспечения заключается в различии окружений. Приложение, которое корректно работает на компьютере разработчика, может вести себя иначе на тестовом или промышленном сервере из-за другой версии библиотеки, отличающихся параметров операционной системы или отсутствия необходимой зависимости. Чем сложнее система и чем больше в ней компонентов, тем острее эта проблема.

Первым решением стала аппаратная виртуализация, при которой на одном физическом сервере запускается несколько виртуальных машин, каждая из которых включает собственную операционную систему. Этот подход обеспечивает сильную изоляцию и остаётся основой большинства облачных сервисов типа IaaS. Вместе с тем полноценная операционная система в каждой виртуальной машине увеличивает объём образа, время запуска и потребление ресурсов, а потому неудобна для частого развёртывания небольших компонентов.

Контейнеризация предлагает иной уровень изоляции: вместо виртуализации оборудования изолируются процессы внутри одной операционной системы. Контейнеры используют общее ядро хоста, поэтому запускаются быстрее и требуют меньше ресурсов [3; 4]. Контейнер можно рассматривать как изолированный процесс, который включает всё необходимое для работы компонента приложения и не зависит от состояния хост-системы [5].

Технические основы контейнеризации

Контейнеры в Linux опираются на два механизма ядра. Пространства имён (namespaces) ограничивают то, что процесс «видит»: собственные идентификаторы процессов, сетевые интерфейсы, файловую систему и имена узлов. Контрольные группы (cgroups) ограничивают то, сколько ресурсов процесс может использовать: процессорное время, оперативную память, дисковый и сетевой ввод-вывод [3]. Сочетание этих механизмов создаёт для приложения иллюзию отдельной системы при отсутствии накладных расходов полноценной виртуализации.

Для распространения контейнеров используется понятие образа – неизменяемого пакета, в который входят код приложения, среда выполнения, системные библиотеки и настройки. Образ строится из слоёв: каждая инструкция в описании сборки добавляет новый слой, что позволяет повторно использовать неизменённые части и экономить место и время передачи по сети. Готовые образы хранятся в реестрах, откуда их загружают на любой узел, способный запускать контейнеры. Для обеспечения совместимости инструментов разных производителей разработаны отраслевые спецификации формата образов и среды выполнения, поддерживаемые проектом Open Container Initiative [6].

Основные различия между виртуальными машинами и контейнерами представлены в таблице 1.

Таблица 1

Сравнение виртуальных машин и контейнеров

Критерий	**Виртуальная машина**	**Контейнер**
Ядро ОС	Собственное ядро гостевой ОС	Общее ядро хостовой ОС
Размер	Как правило, гигабайты (полный образ ОС)	Как правило, десятки–сотни мегабайт
Время запуска	От десятков секунд до минут	От долей секунды до нескольких секунд

Критерий	Виртуальная машина	Контейнер
Изоляция	Более строгая, на уровне гипервизора	На уровне процессов, слабее
Плотность размещения	Ниже	Выше
Типичное назначение	Различные ОС, строгая изоляция арендаторов	Упаковка и запуск компонентов приложения

Из сравнения следует, что контейнеры не заменяют виртуальные машины, а дополняют их. В облачной среде контейнеры чаще всего запускаются именно внутри виртуальных машин, что позволяет сочетать сильную изоляцию на уровне инфраструктуры с высокой гибкостью на уровне приложений.

Контейнеры и микросервисная архитектура

Контейнеры тесно связаны с микросервисным подходом к построению систем. Согласно классическому определению, микросервисная архитектура представляет собой способ разработки приложения как набора небольших сервисов, каждый из которых выполняется в собственном процессе и взаимодействует с другими через легковесные механизмы, чаще всего через HTTP API [7]. Каждый сервис отвечает за отдельную бизнес-возможность и может разрабатываться, развёртываться и масштабироваться независимо от остальных.

Контейнер является естественной единицей развёртывания микросервиса: он содержит сервис вместе с его зависимостями и может быть запущен в любом количестве экземпляров. Это позволяет, например, разным командам использовать разные языки программирования и версии библиотек без конфликтов между собой, а также обновлять один сервис, не затрагивая остальные [8].

Однако микросервисный подход не является универсальным решением. Он превращает внутренние вызовы функций в сетевые взаимодействия, что увеличивает задержки и число возможных отказов, усложняет обеспечение согласованности данных, отладку и наблюдаемость системы [8; 9]. Для небольших приложений с ограниченной командой разработки монолитная архитектура может оказаться проще и экономичнее.

Оркестрация контейнеров и платформа Kubernetes

Запуск нескольких контейнеров вручную допустим для разработки, но не для промышленной эксплуатации. Когда в системе десятки сервисов и сотни экземпляров контейнеров, необходимо автоматически решать вопросы размещения контейнеров по узлам, перезапуска при отказах, распределения нагрузки, обновления версий и масштабирования. Совокупность таких задач называется оркестрацией.

Kubernetes был создан в Google на основе многолетнего опыта эксплуатации внутренних систем управления кластерами, прежде всего системы Borg [10], и впоследствии передан в фонд Cloud Native Computing Foundation [11]. Платформа предназначена для автоматизации развёртывания, масштабирования и управления контейнеризированными приложениями [12].

Ключевой принцип Kubernetes – декларативное управление. Администратор или разработчик не выполняет последовательность команд, а описывает желаемое состояние системы: какой образ должен использоваться, сколько экземпляров должно работать, какие ресурсы им выделены. Управляющие компоненты платформы непрерывно сравнивают фактическое состояние с желаемым и при расхождении выполняют необходимые действия, например запускают новый экземпляр вместо вышедшего из строя [11; 12].

В основе модели лежат несколько базовых объектов. Под (Pod) – минимальная единица развёртывания, включающая один или несколько тесно связанных контейнеров с общей сетью и хранилищем. Объект Deployment описывает нужное количество экземпляров пода и стратегию обновления, позволяя выполнять постепенное обновление без остановки сервиса и откат к предыдущей версии. Объект Service предоставляет для набора подов постоянный сетевой адрес и распределяет между ними трафик. Автоматическое масштабирование позволяет изменять число экземпляров в зависимости от нагрузки, например по загрузке процессора [12; 13].

Таким образом, Kubernetes выступает своего рода операционной системой для кластера: разработчик взаимодействует с единым программным интерфейсом, а вопросы распределения ресурсов между конкретными серверами скрываются от него. Для контроля состояния такой системы дополнительно необходимы средства наблюдаемости: сбор журналов работы сервисов, метрик нагрузки и отказов, а также трассировка запросов между сервисами. Без них диагностика ошибок в среде из десятков динамически создаваемых и удаляемых контейнеров становится крайне затруднительной.

Роль контейнеров в cloud-native подходе, DevOps и CI/CD

Определение cloud-native, предложенное фондом CNCF, описывает подход к созданию масштабируемых приложений в современных динамических средах, к которым относятся публичные, частные и гибридные облака. В качестве характерных признаков названы контейнеры, сервисные сетки, микросервисы, неизменяемая инфраструктура и декларативные интерфейсы программирования [14]. Контейнеризация в этом перечне играет базовую роль, так как именно она делает приложение переносимым и стандартизированным.

Контейнеры хорошо сочетаются с практиками DevOps, направленными на объединение разработки и эксплуатации и автоматизацию всего пути программного изменения от исходного кода до промышленной среды. Исследования эффективности процессов разработки, обобщённые в работе [15], связывают высокие показатели поставки программного обеспечения с автоматизацией сборки, тестирования и развёртывания, а также с быстрой обратной связью.

Типовой конвейер непрерывной интеграции и доставки (CI/CD) для контейнерного приложения включает следующие этапы: разработчик вносит изменение в репозиторий; система автоматически собирает образ и запускает тесты; успешно проверенный образ помещается в реестр; описание развёртывания в Kubernetes обновляется, и платформа постепенно заменяет старые экземпляры новыми. Если после обновления обнаружены проблемы, возможен быстрый откат к предыдущей версии образа. Дополнительно конфигурация инфраструктуры и приложения описывается в виде кода и хранится в системе контроля версий, что делает изменения прозрачными, воспроизводимыми и проверяемыми [15; 16].

Безопасность и ограничения контейнерных технологий

Несмотря на очевидные преимущества, контейнерные платформы создают собственные риски. Национальный институт стандартов и технологий США в специальной публикации по безопасности контейнеров выделяет уязвимости образов, небезопасную настройку реестров, риски оркестратора, риски самих контейнеров и риски хостовой операционной системы [17]. Поскольку контейнеры используют общее ядро, уязвимость ядра или ошибка в настройке изоляции может затронуть все контейнеры на узле, тогда как изоляция виртуальных машин обычно строже.

На практике снижение рисков достигается комплексом мер: использованием проверенных и минимальных базовых образов, регулярным сканированием образов на известные уязвимости, запретом запуска контейнеров с избыточными привилегиями,

безопасным хранением секретов, ограничением сетевого взаимодействия между сервисами и разграничением прав доступа к самому оркестратору [17].

Помимо безопасности, следует учитывать другие ограничения. Kubernetes сам по себе является сложной системой, требующей квалификации для настройки, сопровождения и диагностики. Приложения с сохранением состояния, такие как базы данных, сложнее запускать в контейнерах, чем не хранящие состояние сервисы, поэтому базы данных нередко размещают вне кластера, используя управляемые облачные сервисы. Кроме того, для небольших проектов затраты на создание и поддержку контейнерной платформы могут превысить получаемую пользу; в таких случаях оправданно использование более простых решений, например виртуальной машины или управляемой платформы PaaS.

Пример: контейнерная архитектура интернет-магазина

Рассмотрим применение описанных технологий на примере интернет-магазина, состоящего из веб-интерфейса и нескольких серверных сервисов: каталога товаров, корзины, обработки заказов, платежей и уведомлений. Схема такой архитектуры представлена на рисунке 1. Каждый сервис упаковывается в отдельный образ и описывается собственным объектом Deployment, а доступ между сервисами и извне обеспечивается объектами Service и точкой входа для внешнего трафика (Ingress).

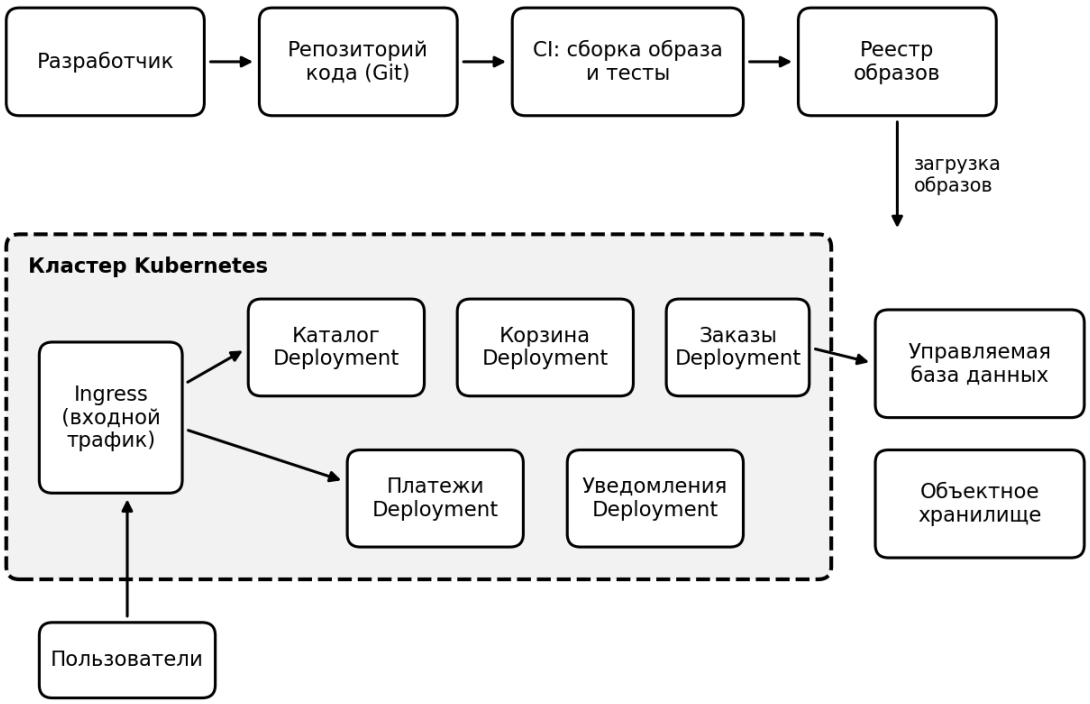


Рисунок 1 – Схема контейнерной архитектуры интернет-магазина

Нагрузка на такую систему неравномерна. Во время рекламной акции резко возрастает число обращений к каталогу и корзине, тогда как нагрузка на модуль уведомлений меняется незначительно. Благодаря независимому масштабированию платформа может увеличить число экземпляров только перегруженных сервисов, не расходуя ресурсы на остальные, а после спада нагрузки – вернуть прежнее количество. При сбое одного из узлов кластера оркестратор автоматически запустит потерянные экземпляры на исправных узлах.

Базу данных заказов и покупателей целесообразно разместить в управляемом облачном сервисе с резервным копированием, а изображения товаров хранить в объектном хранилище. Обновление сервисов выполняется через конвейер CI/CD: при выходе новой версии платёжного модуля постепенно заменяются только его экземпляры,

а в случае ошибки происходит откат. Такой подход сокращает время выпуска изменений и снижает риск простоя магазина при обновлениях.

Рекомендации по внедрению контейнерной платформы

Внедрение контейнеризации целесообразно начинать не с выбора инструментов, а с анализа требований проекта: ожидаемой нагрузки, требований к доступности, структуры приложения, квалификации команды и бюджета. На основе практики, описанной в источниках [16; 17], можно сформулировать следующие рекомендации.

Во-первых, переход стоит выполнять постепенно: сначала упаковать в контейнеры существующие компоненты и настроить автоматическую сборку образов, затем перейти к оркестрации и лишь при наличии реальной необходимости – к разделению приложения на микросервисы. Во-вторых, требования безопасности следует закладывать с самого начала, включив сканирование образов и проверку конфигураций в конвейер CI/CD. В-третьих, необходимо заранее внедрить мониторинг, централизованное журналирование и процедуры восстановления после сбоев. В-четвёртых, для сервисов с сохранением состояния следует оценить возможность использования управляемых облачных сервисов вместо самостоятельной эксплуатации в кластере. Наконец, для небольших проектов имеет смысл сравнить затраты на собственный кластер с использованием управляемых решений, в которых часть задач эксплуатации берёт на себя поставщик.

Заключение

Контейнеризация и оркестрация стали технологической основой cloud-native разработки, так как решают ключевую проблему облачных систем – единообразную упаковку, доставку и автоматическое управление программным обеспечением. Контейнеры обеспечивают переносимость и воспроизводимость среды выполнения, а системы оркестрации, прежде всего Kubernetes, позволяют управлять большим числом сервисов на основе декларативного описания желаемого состояния.

Показано, что контейнеры не заменяют виртуальные машины, а работают поверх них, сочетая изоляцию инфраструктуры и гибкость приложений. Совместное использование контейнеров, микросервисной архитектуры и практик CI/CD сокращает время выпуска изменений и повышает устойчивость систем, однако увеличивает сложность архитектуры и требует внимания к безопасности образов, реестров, оркестратора и хостовых систем.

Выбор контейнерной платформы должен определяться требованиями конкретного проекта: масштабом нагрузки, структурой приложения, квалификацией команды и допустимыми затратами. Дальнейшее развитие технологий связано с автоматизацией управления инфраструктурой, усилением встроенных механизмов безопасности и применением контейнерных платформ для распределённых вычислений, в том числе для задач машинного обучения и периферийных вычислений.

Список литературы

1. *Мелл П., Гранс Т.* Определение облачных вычислений NIST [на англ. яз.]. – Гейтерсберг: Национальный институт стандартов и технологий США, 2011. – Специальная публикация NIST 800-145. – DOI: 10.6028/NIST.SP.800-145.
2. *Меркель Д.* Docker: легковесные контейнеры Linux для согласованной разработки и развёртывания [на англ. яз.] // Linux Journal. – 2014. – Вып. 239. – Ст. 2.
3. *Бернстайн Д.* Контейнеры и облако: от LXC к Docker и Kubernetes [на англ. яз.] // IEEE Cloud Computing. – 2014. – Т. 1, № 3. – С. 81–84. – DOI: 10.1109/MCC.2014.51.
4. *Пал К.* Контейнеризация и облачная платформа PaaS [на англ. яз.] // IEEE Cloud Computing. – 2015. – Т. 2, № 3. – С. 24–31. – DOI: 10.1109/MCC.2015.51.
5. *Миелл И., Сейерс Э.* Docker на практике. – М.: ДМК Пресс, 2020.

6. *Инициатива открытых контейнеров.* Спецификации формата образов и среды выполнения контейнеров [на англ. яз.]. – URL: https://opencontainers.org/
7. *Льюис Дж., Фаулер М.* Микросервисы: определение нового архитектурного термина [на англ. яз.]. – 2014. – URL: https://martinfowler.com/articles/microservices.html
8. *Драгони Н., Джалоренцо С., Лафуэнте А. Л., Маццара М., Монтези Ф., Мустафин Р., Сафина Л.* Микросервисы: вчера, сегодня и завтра [на англ. яз.] // Present and Ulterior Software Engineering. – Cham: Springer, 2017. – С. 195–216. – DOI: 10.1007/978-3-319-67425-4_12.
9. *Ричардсон К.* Микросервисы. Паттерны разработки и рефакторинга. – СПб.: Питер, 2019.
10. *Верма А., Педроса Л., Корупулу М., Оппенхаймер Д., Тьюн Э., Уилкс Дж.* Управление кластерами в крупном масштабе в Google с помощью Borg [на англ. яз.] // Материалы 10-й Европейской конференции по компьютерным системам (EuroSys '15). – Нью-Йорк: ACM, 2015. – Ст. 18. – DOI: 10.1145/2741948.2741964.
11. *Бёрнс Б., Грант Б., Оппенхаймер Д., Брюер Э., Уилкс Дж.* Borg, Omega и Kubernetes [на англ. яз.] // ACM Queue. – 2016. – Т. 14, № 1. – С. 70–93.
12. *Авторы проекта Kubernetes.* Что такое Kubernetes? – Документация Kubernetes. – URL: https://kubernetes.io/ru/docs/concepts/overview/what-is-kubernetes/
13. *Лукша М.* Kubernetes в действии. – М.: ДМК Пресс, 2019.
14. *Cloud Native Computing Foundation.* Определение Cloud Native, версия 1.0 [на англ. яз.]. – URL: https://github.com/cncf/toc/blob/main/DEFINITION.md
15. *Форсгрен Н., Хамбл Дж., Ким Дж.* Ускоряйся! Наука о бережливой разработке программного обеспечения и DevOps [на англ. яз.]. – Портленд: IT Revolution Press, 2018.
16. *Ким Дж., Хамбл Дж., Дебуа П., Уиллис Дж.* Руководство по DevOps. – М.: Манн, Иванов и Фербер, 2019.
17. *Супайя М., Морелло Дж., Скарфоун К.* Руководство по безопасности контейнеров приложений [на англ. яз.]. – Гейтерсберг: Национальный институт стандартов и технологий США, 2017. – Специальная публикация NIST 800-190. – DOI: 10.6028/NIST.SP.800-190.

FLUX-128 ЖЕҢІЛСАЛМАҚТЫ КРИПТОГРАФИЯЛЫҚ ХЕШ-ФУНКЦИЯСЫНЫҢ ЛАВИНДЫҚ СИПАТТАМАЛАРЫН ЗЕРТТЕУ

Алғазы К.Т., Әлімжан Е.Ж., Сақан Қ.С.
e-mail: kunbolat@mail.ru, ayerkebulan19@gmail.com, 19kairat78@gmail.com
ҚР ҒК ҒЖБМ Ақпараттық және есептеу технологиялары институты, Қазақстан

Аңдатпа. *Мақалада Заттар интернеті жүйелерінде, ендірілген құрылғыларда және есептеу ресурстары шектеулі басқа да платформаларда қолдануға арналған жаңа жеңіл криптографиялық хеш-функцияның құрылымы ұсынылған. Сондай-ақ оның диффузиялық қасиеттерін эксперименттік зерттеу нәтижелері келтірілген. Алгоритмді бағалау қатаң лавиндық критерийді (SAC) талдауға негізделген, ол хеш-мәннің әрбір шығыс битінің жекелеген кіріс биттерінің инверсиясына сезімталдығын сандық тұрғыдан бағалауға мүмкіндік береді. SAC бағалау үшін ұзындығы 72 бит болатын 5000 түрлі кіріс хабарламасынан тұратын деректер жиыны пайдаланылды. Әрбір хабарлама үшін 128 биттік хеш мәні есептелді, содан кейін әрбір кіріс биті кезекпен инверсияланып, хеш қайта есептелді. Бұл процедура «кіріс биті–шығыс биті» жұптары үшін шығыс битінің ауысу ықтималдықтары матрицаларын құруға және байқалған ықтималдықтардың 0,5 мақсатты мәніне жақындығын бағалауға мүмкіндік берді. Нәтижелер алынған мәндердің позициялар мен раундтар бойынша таралуын көрсететін визуализациялармен толықтырылды, бұл өзгергіштік заңдылықтарын нақты бағалауға мүмкіндік береді. Алынған нәтижелер тереңдетілген криптоталдауды және алгоритмнің коллизияларға төзімділігін бағалауды қоса алғанда, кейінгі зерттеулерге негіз болады.*

1 Кіріспе

Заттар интернетінің, ендірілген жүйелердің және ресурстары шектеулі басқа да платформалардың дамуы практикалық қолдану сценарийлерінде қолайлы өнімділік пен жеткілікті беріктік деңгейін бір мезгілде қамтамасыз ететін криптографиялық примитивтерге сұраныс қалыптастырды. Осыған байланысты симметриялық алгоритмдерді кешенді бағалау әдістеріне назар күшейді, мұнда криптоталдаумен қатар түрлендіру сапасының инженерлік және статистикалық көрсеткіштері қарастырылады. Атап айтқанда, АҚШ Ұлттық стандарттар және технологиялар институтының (National Institute of Standards and Technology, NIST) жеңіл криптографияны стандарттау процесі аясында кандидаттарды жан-жақты талдау және нәтижелердің салыстырмалылығын қамтамасыз ету қажеттілігі, оның ішінде диффузия қасиеттері мен қолданбалы пайдалану режимдеріндегі шығыс тізбектерінің мінез-құлқын бағалау қажеттілігі атап өтіледі [1].

NIST-тің жеңілсалмақты криптография саласындағы тәжірибесі ресурстары шектеулі құрылғылар үшін дәстүрлі стандарттар іске асыру шығындары тұрғысынан оңтайлы болмаған жағдайларда балама бола алатын примитивтердің сұранысқа ие екенін көрсетеді. 2025 жылы NIST шектеулі құрылғыларға арналған аутентификацияланған хештеу және шифрлау үшін Ascon тобының стандартын жариялап, техникалық спецификациялар мен қолдану контекстін бекітті [2]. Сонымен қатар қолданбалы криптографияда негізгі салыстыру нүктесі ретінде NIST 2023 жылғы редакциясында алгоритмнің техникалық мазмұнын өзгертпей жаңартқан AES қолданысын жалғастыруда [3]. IoT үшін жеңілсалмақты криптографияға арналған шолу зерттеулері іске асыру құны, жылдамдық және қауіпсіздік арасындағы ымыраларды жүйелейді, сондай-ақ нақты шектеулер жағдайында алгоритмдерді салыстыру үшін түрлендіру сапасының эксперименттік метрикаларының өзектілігін растайды [4, 5].

Қазіргі симметриялық түрлендірулерге қойылатын негізгі талаптардың бірі айқын диффузия (diffusion) болып табылады, мұнда кірістегі шағын өзгеріс шығыста ауқымды өзгерістерге әкеледі. Мұндай мінез-құлықтың кең таралған критерийлерінің бірі – қатаң лавиндық критерий (Strict Avalanche Criterion, SAC), ол бір кіріс биті инверсияланған кезде әрбір шығыс битінің инверсиялану ықтималдығының 0.5 мәніне жақын болуын формалдайды. Мұндай қасиеттер жеңіл блоктық шифрларды жобалау кезінде де, S-box типті компоненттерді құрастыру және талдау кезінде де белсенді қолданылады, өйткені олар ерте раундтардағы араластыру (confusion) сапасын және бірқатар статистикалық әрі дифференциалдық әсерлерге төзімділікті едәуір дәрежеде анықтайды. Бұл «жақсы лавиндық әсер» негізгі пайдалану сипаттамаларының бірі ретінде қарастырылатын жеңіл конструкцияларды зерттеу жұмыстарымен расталады [6]. SAC критерийін қанағаттандыратын S-box синтезі жөніндегі жұмыстарда идеалға жақын лавиндық көрсеткіштерге қол жеткізу бейсызықтықты және құрылымдық тәуелділіктерге төзімділікті қамтамасыз етумен байланысты жеке жобалау міндеті екені атап көрсетіледі [7]. Сонымен қатар хаостық және гибридтік S-box синтездеу тәсілдері жеңілдікті сақтай отырып қажетті лавиндық параметрлерді қамтамасыз ете алатыны көрсетіледі, бұл тақырыпты шектеулі іске асырулар үшін өзекті етеді [8].

Алгоритмнің эксперименттік режимдерде шығыс биттік тізбектердің генераторы ретіндегі мінез-құлқын дұрыс сипаттау үшін диффузияның тек биттік критерийлері жеткіліксіз. Сондықтан практикада NIST STS сияқты статистикалық тестілер жиынтығы жиі қолданылады. Олар кездейсоқ тізбектен күтілетін мінез-құлықтан ауытқулардың әртүрлі кластарының бар-жоғын анықтап, p-values мәндері мен тестілерден өту жиіліктерінің жиынтығын қалыптастырады [9, 10].

2Материалдар мен әдістер

Хештеу алгоритмі үшін қатаң лавиндық критерий келесі өрнекпен анықталады:

$$\varepsilon_s = \left|2p_{s,i,j} - 1\right|,$$

мұндағы, $i = 1, \dots, 72$– талданатын шығыс битінің нөмірі, $j = 1, \dots, 128$– инверсияланатын кіріс битінің нөмірі, ал $p_{s,i,j}$– i-ші шығыс биті j-ші кіріс биті инверсияланған кезде өзгеретінінің эмпирикалық ықтималдығы.

Критерий барлық (i, j) жұптары үшін $p_{s,i,j}$ мәндері 0.5 шамасына жақын болса, ал ε_s модулі бойынша аз болса, орындалды деп есептеледі.

Эксперимент келесідей ұйымдастырылады. 5000 түрлі бастапқы хабарламадан тұратын жиын қарастырылады P_k^0, $k = 1, 2, \dots, 5000$.

Әрбір P_k^0 хабарламасы үшін C_k^0 деп белгіленетін базалық хеш мәні есептеледі. Одан әрі бекітілген k үшін нөмірі $j = 1, \dots, 128$ болатын әрбір кіріс биті кезекпен инверсияланады, нәтижесінде өзгертілген $P_k^{(j)}$ хабарламалары қалыптасады, олар үшін $C_k^{(j)}$ хеш-мәндері қайта есептеледі.

Содан кейін әрбір (i, j) индекстер жұбы үшін i-ші хеш биті $C_k^{(j)}$ бастапқы хештің i-ші битімен C_k^0 салыстырылады. Бекітілген k үшін осы салыстыру нәтижелерін Q_k өлшемі 72×128 болатын матрица түрінде көрсету ыңғайлы:

$$Q_k = \begin{pmatrix} q_{1,1}^k & q_{1,2}^k & \cdots & q_{1,128}^k \\ q_{2,1}^k & q_{2,2}^k & \cdots & q_{2,128}^k \\ \vdots & \vdots & \ddots & \vdots \\ q_{72,1}^k & q_{72,2}^k & \cdots & q_{72,128}^k \end{pmatrix}$$

мұндағы, $q_{i,j}^k = \begin{cases} 1, \text{ егер } i-\text{ші } C_k^j \text{ биті } i-\text{ші } C_k^0 \text{ битінен өзгеше болса;} \\ 0, \text{ егер } i-\text{ші бит өзгермесе.} \end{cases}$

Барлық хабарламалар бойынша $p_{s,i,j}$ ықтималдықтарын бағалау үшін Q_k матрицаларының сәйкес элементтерін k индексі бойынша қосамыз. Осылайша өлшемі R болатын 72×128 матрицасын құрамыз.

Матрицаның әрбір R элементін хабарламалар санына (5000) бөлу арқылы эмпирикалық ықтималдықтар P^s матрицасын аламыз:

$$P^s = \begin{pmatrix} p_{s,1,1} & p_{s,1,2} & \cdots & p_{s,1,128} \\ p_{s,2,1} & p_{s,2,2} & \cdots & p_{s,2,128} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s,72,1} & p_{s,72,2} & \cdots & p_{s,72,128} \end{pmatrix}$$

мұнда,

$$p_{s,i,j} = \frac{1}{5000} \sum_{k=1}^{5000} q_{i,j}^k,\ i = 1,\ \ldots,\ 72,\ j = 1,\ \ldots,\ 128.$$

P^s матрицасының негізінде зерттелетін хеш-функцияның барлық (i, j) жұптары үшін қатаң лавиндық критерийді қанағаттандыруы бағаланады. Практикада бұл үшін таңдалған 5000 хабарлама бойынша Q_k, R және P^s матрицаларын автоматты түрде қалыптастыратын, олардың өлшемі 72×128 болатын және $p_{s,i,j}$ мәндерінің идеал 0.5 деңгейінен ауытқуын талдауға мүмкіндік беретін арнайы бағдарлама әзірленді.

Бұдан әрі SAC талдауынан бөлек, FLUX-128 шығыстарының диффузия және псевдокездейсоқтық қасиеттерін кеңейтілген статистикалық верификациялау үшін NIST STS статистикалық кездейсоқтық тестілерінің жиынтығы қолданылды. Тексеру ұзындығы тұрақты кіріс хабарламалары сериясы үшін 128 биттік хеш-мәндерді конкатенациялау арқылы қалыптастырылған екілік ағын бойынша орындалды.

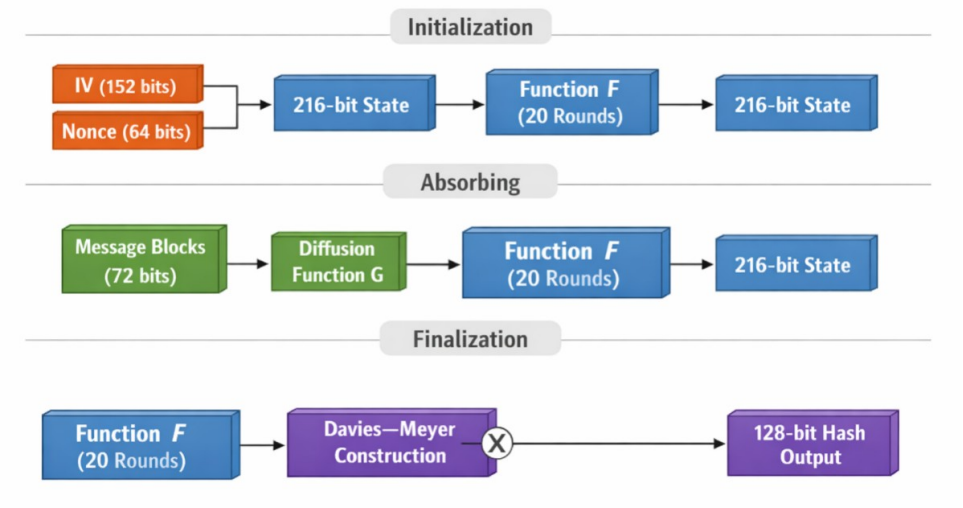
3Нәтижелер және талқылау

3.1 FLUX-128 жеңілсалмақты криптографиялық хеш-функциясы

FLUX-128 – жад, аудан және энергия тұтыну шектеулері маңызды болатын IoT және ендірілген платформаларға бағытталған жеңілсалмақты криптографиялық хеш-функция. Алгоритм sponge (губка) конструкциясына негізделген және хабарламаны итеративті түрде өңдейді: кіріс 72 биттік блоктарға бөлінеді, соңғы бөлік толықтырушы биттермен (padding) толықтырылады, содан кейін блоктар тізбегі ішкі күйге «жұтылады». Жұмыста мына параметрлер бекітілген: ішкі күй 216 бит (27 байт), жұту жылдамдығы r = 72 бит (9 байт) және ұзындығы 128 бит болатын шығыс тегін қалыптастыру.

Бұл алгоритмнің ғылыми жаңалығы IoT ортасының шектеулеріне бейімделген дәстүрлі sponge конструкциясын модификациялауда. Базалық F түрлендіруінің алдында қосымша жеңілдетілген G кезеңі енгізілетін сұлба ұсынылады. Бұл функция кішірейтілген ішкі күймен жұмыс істейді және жоғары өткізу қабілетіне есептелген, соның арқасында мақсатты криптографиялық құрылымды сақтай отырып деректерді өңдеу жеделдетіледі.

Хешті есептеу үш кезеңнен тұрады: инициализациялау, жұту және финализациялау. FLUX-128 алгоритмінің құрылымдық ұйымдастырылуы және оның негізгі кезеңдерінің (инициализациялау, хабарламаларды жұту және финализациялау) өзара байланысы 1-суретте көрсетілген, ол ішкі күй түрлендірулерінің ретін бейнелейді.



Сурет 1 – FLUX-128 хештеу алгоритмінің құрылымдық сұлбасы

Инициализациялау кезеңінде 216 биттік күй 152 биттік IV және 64 биттік nonce мәндерінен құралады, содан кейін R = 20 раунд саны бар F раундтық функциясымен түрлендіріледі. «Жұту» (absorbing) кезеңінде әрбір 72 биттік блок алдымен 3×3 байт матрицасы негізінде іске асырылған және S-box арқылы сызықтық араластыруды әрі 72 биттік ағынның 1 биттік циклдік ротациясын қамтитын көмекші G диффузиялық функциясы (кіріс/шығыс 72 бит) арқылы өтеді, содан кейін алынған нәтиже F көмегімен 216 биттік күйді жаңарту үшін пайдаланылады. Осылайша, G блок ішіндегі жергілікті диффузияны күшейтеді, ал F барлық раундтар бойы блоктар ықпалының күйде жинақталуын қамтамасыз етеді.

Финализациялау 20 раундтық F функциясын қайталап қолдану және Davies–Meyer сұлбасы бойынша араластыру арқылы орындалады, мұнда шығыс биттерінің бір бөлігі финализациялауға дейін күйдің сәйкес биттерімен модуль 2 бойынша (XOR) қосымша қосылады; толық 128 биттік тегті алу үшін «сығу» (squeezing) процедурасы екі рет орындалады. F раундтық функциясы XOR операцияларына, S-box алмастыруларына және байттық орын ауыстыруларға (бекітілген PerF арқылы) негізделеді, бұл қажетті диффузия деңгейін сақтай отырып аппараттық қарапайымдыққа бағытталған мақсатқа сәйкес келеді.

Absorbing кезеңіне қатысатын *G* функциясының жұмыс істеу принциптері, сондай-ақ *F* функциясының Absorbing және Finalization кезеңдеріндегі жұмыс істеу принциптері төменде келтірілген.

***G* функциясы.** Кірістегі 72 биттік (9 байттық) ақпарат D матрицасының элементтері ретінде қарастырылады:

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix}.$$

G функциясының бүкіл түрлендіру жұмысын келесі үш процедура арқылы жазуға болады: G_1, G_2 және G_3. G_1 және G_3 процедураларының ерекшелігі – өңдеу барысында сызықтық және бейсызық криптографиялық тораптардың ретімен қолданылуы. G_1 процедурасы берілген *D* матрицасынан жаңа матрица алу үшін қолданылады және оны (1) формуламен сипаттауға болады:

$$\left.\begin{array}{c} q = \sum_{k=1}^{3} d_{ik} \oplus \left(\sum_{\substack{k=1 \\ k \neq i}}^{3} d_{kj} \right); \\ d_{ij} = S(q); \end{array}\right\} \; i = 1,\ 2,\ 3;\ j = 1,\ 2,\ 3. \tag{1}$$

мұндағы, q – аралық мән, $\sum$ операциясы модуль 2 бойынша қосу операциясын (биттік XOR) білдіреді. S – 1-кестеге сәйкес жарты байттар деңгейінде орындалатын алмастыру кестесі (S-блок).

Кесте 1 – S-блок

x	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
S	2	E	F	5	C	1	9	A	B	4	6	8	0	7	3	D

G_2 процедурасы бір битке солға циклдік ығыстыру операциясын білдіреді: $D = D \lll 1$. G_3 процедурасының құрылымы жоғарыда сипатталған G_1, процедурасына ұқсас, айырмашылығы жаңа матрица элементтерін есептеу тәртібінде. (2) формула G_3 әрекетін анықтайды:

$$\left.\begin{array}{c} q = \bigoplus \sum_{k=1}^{3} d_{ik} \bigoplus \left(\bigoplus \sum_{\substack{k=1 \\ k \neq i}}^{3} d_{kj} \right); \\ d_{ij} = S(q); \end{array}\right\} \; i = 3,\ 2,\ 1;\ \ j = 3,\ 2,\ 1. \tag{2}$$

G функциясының жұмысының нәтижесі – жаңартылған байттық D матрицасы.

Раундтық функция *F*, *F₁, F₂ және F₃,* ішкі функцияларынан тұрады және алгоритмнің барлық кезеңдерінде 20 раунд орындайды. Ерекшелік – *F₁*: ол тек бірінші раундта қолданылады, ал екінші раундтан соңғы раундқа дейін раундтық функция *F* құрамына тек *F₂ және F₃* кіреді.*F* раундтық функциясының күй ұзындығы 216 бит немесе 27 байт, олар әрқайсысы 3×3 өлшемді үш квадрат матрицаға ретімен бөлінеді: *A, B* және *C*.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}, \quad C = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix}.$$

Раундтық *F* функциясының жұмысын келесі үш ішкі функция түрінде қарастыруға болады: *F₁, F₂* және *F₃*.

Кіріс мәні **ішкі функция *F₁*** үшін *G* функциясының нәтижесі – *D* матрицасы. Кіріс *D* матрицасының әр жолындағы үш байт сәйкесінше *A, B* және *C* матрицаларының негізгі диагоналіндегі әрбір элементпен XOR бойынша элементтік түрде қосылады. Алынған қосындылар S-блок арқылы өткізіледі және олардың негізінде жаңа матрицалар қалыптастырылады: *A, B* және *C*.

Ішкі функция *F₂* A, B және C матрицаларының жаңа диагональ элементтерін ($a_{ii},\ b_{ii},\ c_{ii},\ i = 1, 2, 3$) (3) формула бойынша анықтайды.

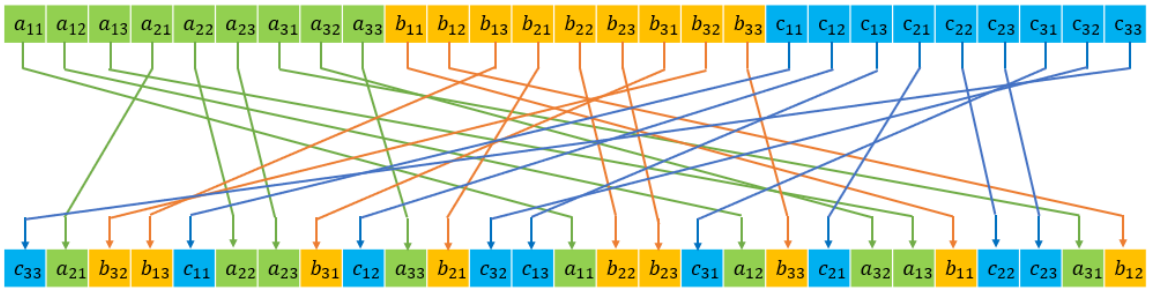
$$\left.\begin{array}{c} q\ = \sum_{j=1}^{3} x_{ij} \bigoplus \sum_{\substack{j=1 \\ j \neq i}}^{3} x_{ji}; \\ x_{ii} = S(q\ ); \end{array}\right\} \; i = 1,\ 2,\ 3; \tag{3}$$

мұндағы x_{ij} – матрицалардың элементтері: *A, B* және *C*.

Ішкі функция $\mathbf{F_{3\text{-те}}}$ матрицалардың мәндері *A, B* және *C PerF* байттық орын ауыстыру түрлендіруіне ұшырайды. Матрицалардың жаңа мәндері келесідей анықталады:

$$A = \begin{pmatrix} c_{33} & a_{21} & b_{32} \\ b_{13} & c_{11} & a_{22} \\ a_{23} & b_{31} & c_{12} \end{pmatrix}, \quad B = \begin{pmatrix} a_{33} & b_{21} & c_{32} \\ c_{13} & a_{11} & b_{22} \\ b_{23} & c_{31} & a_{12} \end{pmatrix}, \quad C = \begin{pmatrix} b_{33} & c_{21} & a_{32} \\ a_{13} & b_{11} & c_{22} \\ c_{23} & a_{31} & b_{12} \end{pmatrix}.$$

Төменде 2-сурет *PerF* орын ауыстыру түрлендіруін графикалық түрде сипаттайды:



Сурет 2 - PerF түрлендірулері

Ұсынылған FLUX-128 хеш-функциясының өнімділігі бағдарламалық салыстырмалы талдау әдістемесі арқылы бағаланды. Бағалау негізгі тиімділік метрикалары бойынша жүргізілді, оның ішінде бір байтқа шаққандағы циклдер саны (CPB), жедел жадты (RAM) және тұрақты жадты (ROM) пайдалану, сондай-ақ бір байтқа шаққандағы энергия тұтыну (Кесте 2). ChipWhisperer сияқты аппараттық өлшеу құралдары қолжетімсіз болғандықтан, циклдер саны элементар операциялар саны мен AVR (Advanced Virtual RISC) командаларының уақыттық моделіне сүйене отырып аналитикалық түрде бағаланды. Энергия тұтыну 8 МГц жиілікте V = 5 В және I = 8 мА деп қабылдайтын AVR қуат моделінің көмегімен есептелді. Жадқа қойылатын талаптар алгоритм құрылымы мен кодтың есептік көлемінен алынды. Салыстырмалы бағалау үшін SHA-256, PHOTON, BLAKE3 және ASCON тестілерінің жарияланған нәтижелері пайдаланылды [11].

Кесте 2 - 8 биттік AVR платформасындағы өнімділікті салыстыру

Хеш алгоритмдері	CPB	RAM(B)	ROM(B)	Энергия(nJ)
PHOTON	845	588	2708	363.6
Ascon	2531	153	2648	1099.3
BLAKE3	2545	3827	48072	1128.06
FLUX-128	681	4228	304	650.06

Эксперименттік талдау нәтижелері ұсынылған FLUX-128 хеш-функциясының 8 биттік AVR платформаларында жоғары есептеу тиімділігіне ие екенін және PHOTON, Ascon-Hash және BLAKE3-пен салыстырғанда cycles/byte көрсеткішінің ең төмен мәнін қамтамасыз ететінін көрсетеді. Сонымен қатар FLUX-128 тұрақты жадты (ROM) өте аз тұтынуымен сипатталады, бұл оны ресурстары шектеулі құрылғылар үшін әсіресе перспективалы етеді.

3.2 SAC нәтижелері

FLUX-128 хеш-алгоритмінің қатаң лавиндық әсерін бағалау үшін ұзындығы 72 бит болатын 5000 түрлі кіріс хабарламасынан тұратын іріктемеде эксперименттер сериясы жүргізілді. Әрбір хабарлама үшін алдымен ұзындығы 128 бит болатын бастапқы хеш-мән есептелді, содан кейін әрбір кіріс биті кезекпен инверсияланып, хеш қайта есептелді. Бұдан кейін бастапқы және өзгертілген шығыстар салыстырылып, берілген кіріс биті инверсияланған кезде нақты шығыс битінің өзгерген-өзгермегені тіркелді.

Осы эксперименттердің негізінде 72×128 өлшемді ықтималдықтар матрицасы қалыптастырылды және ол 3-суретте көрсетілген, мұнда әрбір ұяшық нақты кіріс биті инверсияланған кезде берілген шығыс битінің қаншалықты жиі өзгеретінін көрсетеді.

$$
\begin{array}{cccccccccc}
0.42_{1,1} & 0.51_{1,2} & 0.45_{1,3} & 0.50_{1,4} & 0.54_{1,5} & 0.47_{1,6} & 0.57_{1,7} & \dots & 0.50_{1,128} \\
0.49_{2,1} & 0.56_{2,2} & 0.50_{2,3} & 0.52_{2,4} & 0.51_{2,5} & 0.48_{2,6} & 0.39_{2,7} & \dots & 0.50_{2,128} \\
0.52_{3,1} & 0.49_{3,2} & 0.61_{3,3} & 0.44_{3,4} & 0.43_{3,5} & 0.47_{3,6} & 0.47_{3,7} & \dots & 0.51_{3,128} \\
0.53_{4,1} & 0.52_{4,2} & 0.48_{4,3} & 0.49_{4,4} & 0.55_{4,5} & 0.48_{4,6} & 0.52_{4,7} & \dots & 0.63_{4,128} \\
0.52_{5,1} & 0.55_{5,2} & 0.47_{5,3} & 0.50_{5,4} & 0.49_{5,5} & 0.51_{5,6} & 0.41_{5,7} & \dots & 0.39_{5,128} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0.48_{72,1} & 0.47_{72,2} & 0.48_{72,3} & 0.48_{72,4} & 0.46_{72,5} & 0.55_{72,6} & 0.51_{72,7} & \dots & 0.49_{72,128}
\end{array}
$$

Сурет 3 – 1-раунд үшін лавиндық эксперимент барысында алынған 72 × 128 өлшемді ықтималдықтар матрицасы

Бұл мәндер FLUX-128 алгоритмінің лавиндық мінез-құлқының эмпирикалық бағасы ретінде түсіндіріледі. Қатаң лавиндық критерийге сәйкес келетін идеал жағдайда әрбір кіріс битінің инверсиясы әрбір шығыс битінің 0.5 ықтималдықпен өзгеруіне әкелуі тиіс.

Алынған матрицаны әрі қарай талдау FLUX-128 алгоритміндегі диффузияның біркелкілігін бағалауға, идеал мінез-құлықтан елеулі ауытқулары бар позицияларды анықтауға және қатаң лавиндық критерийдің орындалу дәрежесі туралы қорытынды жасауға мүмкіндік береді. Келесі бөлімде эксперименттің сандық нәтижелері және олардың түсіндірмесі келтіріледі.

Матрица элементтерінің таралуы да алгоритм мінез-құлқының қатаң лавиндық критерий талаптарына жақындығын растайды. Әрбір раунд үшін ықтималдық мәндерінің шамамен 96%-ы (0.40, 0.60) аралығына түседі, яғни шығыс биттерінің көпшілігі кіріс биттерінің инверсиясына идеал деңгейден ±0.10-нан аспайтын ауытқумен жауап береді. Неғұрлым тар 0.45–0.55 интервалы барлық элементтердің шамамен 73%-ын қамтиды, ал одан да тар 0.48–0.52 диапазонында мәндердің шамамен 38%-ы шоғырланған. Позициялардың елеулі бөлігі 0.5 мәніне қатысты іс жүзінде симметриялы мінез-құлық көрсетеді.

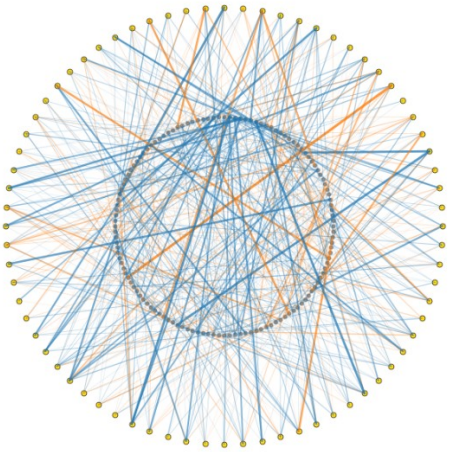
Сонымен бірге матрицаларда жергілікті ауытқулар да байқалады. Әртүрлі раундтар үшін ықтималдықтардың ең төмен мәндері 0.27–0.33 деңгейінде, ал ең жоғары мәндері шамамен 0.66–0.73 деңгейінде орналасқан. Бұл ретте қатты ығысқан элементтердің үлесі аз: 0.5-тен 0.15-тен артық ерекшеленетін мәндер (яғни 0.35–0.65 интервалының сыртында жатқан мәндер) бір пайыздан аз, ал 0.5-тен 0.2-ден артық ауытқитын ықтималдықтар жекелеген жағдайларда ғана кездеседі.

Ықтималдықтар таралуының сипаты раундтан раундқа іс жүзінде өзгермейтінін атап өткен жөн. Орташа мәндер, диапазондар және көрсетілген интервалдардағы элементтердің үлестері барлық 20 раунд үшін бір деңгейде қалады, нашарлау немесе жақсару бағытындағы айқын үрдіс байқалмайды. Бұл алгоритмнің лавиндық қасиеттері салыстырмалы түрде аз раунд санының өзінде қанығатынын, ал раундтық құрылым тереңдігін одан әрі арттыру жаһандық көрсеткіштерге әсер етпей, жергілікті ауытқуларды ғана аздап қайта бөлетінін көрсетеді.

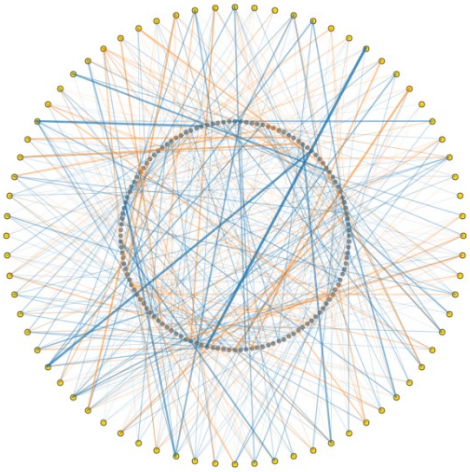
Жалпы алынған нәтижелер 5000 хабарламадан тұратын іріктемеде FLUX-128 алгоритмі қатаң лавиндық критерийдің орындалуына жақын мінез-құлық көрсетеді деген қорытынды жасауға мүмкіндік береді. «Кіріс биті – шығыс биті» жұптарының басым көпшілігі үшін кіріс биті инверсияланған кезде шығыс битінің өзгеру ықтималдығы 0.5 маңында орналасады, ал айқын жүйелі ауытқулар анықталған жоқ.

Қосымша түрде үш сипаттамалық раунд (1-ші, 10-шы және 20-шы) үшін 4–6-суреттерде көрсетілген лавиндық ауытқу диаграммалары (avalanche fingerprints) құрылды. Бұл диаграммаларда ішкі сақинада 72 кіріс биті, сыртқы сақинада 128 шығыс биті орналасқан, ал сызықтармен ықтималдықтың 0.5 мәнінен ең үлкен ауытқуы бар «кіріс–шығыс» жұптары белгіленген. Мұндай визуализация тәсілі идеал лавиндық мінез-

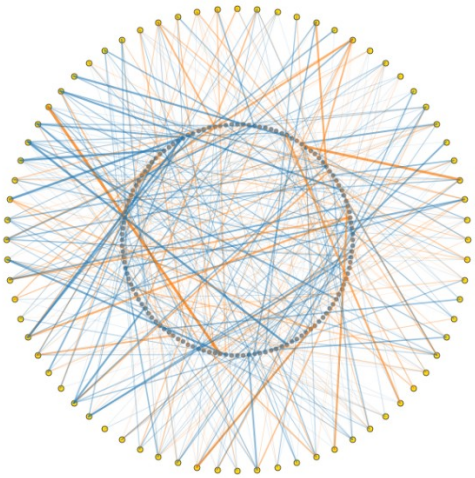
құлықтан ең көп ерекшеленетін тәуелділіктерді көрнекі түрде бөліп көрсетуге мүмкіндік береді.



Сурет 4 – 1-раунд үшін FLUX-128 хеш-функциясының лавиндық ауытқулар диаграммасы



Сурет 5 – 10-раунд үшін FLUX-128 хеш-функциясының лавиндық ауытқулар диаграммасы



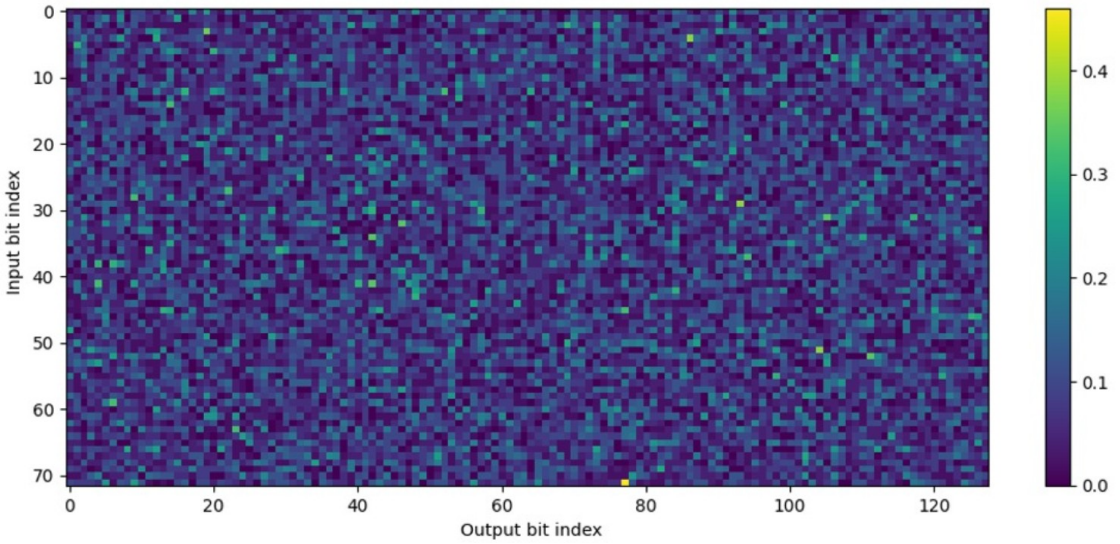
Сурет 6 – 20-раунд үшін FLUX-128 хеш-функциясының лавиндық ауытқулар диаграммасы

Лавиндық ауытқулар диаграммасы лавиндық әсер тестінің нәтижелерін дөңгелек сұлба түрінде көрсетеді, онда әрбір битке шеңбер бойындағы бекітілген позиция сәйкес келеді, ал сол позициядағы визуалды элемент идеал лавиндық мінез-құлықтан ауытқуды бейнелейді. Түс нормаға қатысты ауытқу белгісін кодтайды: көк түс теріс ығысу үшін (мән 0.5-тен төмен), ал қызғылт сары түс оң ығысу үшін (мән 0.5-тен жоғары) қолданылады, бұл теңгерімсіздіктің бағытын бірден анықтауға мүмкіндік береді. Сызық қалыңдығы осы ығысудың шамасын сипаттайды: жіңішке бөліктер шағын ауытқулар мен нормаға жақындыққа сәйкес келеді, ал қалыңдаулар неғұрлым елеулі ауытқуларды көрсетеді. Қалыңдықтың біркелкілігі мен түстер теңгерімі бойынша диффузия тұрақтылығын жылдам бағалап, проблемалық позицияларды анықтауға болады.

Алынған диаграммаларды талдау «күшті» байланыстар саны бит жұптарының жалпы санына қарағанда салыстырмалы түрде аз екенін және мұндай байланыстар кіріс немесе шығыс биттік позициялардың тар жиынында шоғырланбай, диаграмманың бүкіл периметрі бойынша таралғанын көрсетеді. Бұл FLUX-128 алгоритмінде кіріс өзгерістеріне жүйелі түрде басқалардан едәуір нашар жауап беретін айқын «әлсіз» шығыс биттерінің шағын тобы анықталмағанын білдіреді. Керісінше, идеал 0.5 мәнінен елеулі ауытқулар әртүрлі позицияларда пайда болады және тұрақты жергілікті кластерлер қалыптастырмайды.

1-ші, 10-шы және 20-шы раундтардың іздерін салыстыру да раундтар саны артқан сайын алгоритмнің лавиндық қасиеттері тұрақтанатынын растайды. Ерте раундта күшті байланыстар құрылымы неғұрлым ұйымдасқан болып көрінеді: сызықтардың жекелеген топтары визуалды түрде айқынырақ бөлінеді, бұл диффузияның әлі толық қалыптаспағанын көрсетеді. Кейінгі раундтарда таралу біркелкі бола түседі: күшті байланыстар айқын жергіліктенусіз көптеген әртүрлі кіріс және шығыс биттік позицияларды қамтиды. Бұл сандық нәтижелермен сәйкес келеді және орта және кейінгі раундтардың өзінде FLUX-128 алгоритмі қатаң лавиндық критерийдің орындалуына жақын тұрақты мінез-құлық көрсететінін білдіреді.

7-суретте $|\,2p-1\,|$ мәндерінің 20-раунд үшін жылулық картасы көрсетілген, ол кіріс биті инверсияланған кезде шығыс биттерінің өзгеру ықтималдықтары матрицасы бойынша құрылған.



Сурет 7 – $|\,2k-1\,|$ мәндерінің FLUX-128 алгоритмінің 20-раунды үшін жылулық картасы

7-суреттен көрініп тұрғандай, мәндердің таралуы жергілікті «ыстық» аймақтарсыз іс жүзінде біртекті сипатқа ие. Бұл «кіріс биті – шығыс биті» жұптарының басым көпшілігі үшін $|\,2p-1\,|$ шамасы аз болып қалатынын, яғни шығыс битінің байқалатын инверсия ықтималдығы идеал 0.5 деңгейіне жақын екенін білдіреді.

Барлық раундтар және барлық «кіріс–шығыс» жұптары бойынша шығыс биттерінің өзгеру эмпирикалық ықтималдықтары таралуының жинақталған статистикалық сипаттамалары 3-кестеде келтірілген.

Кесте 3 – FLUX-128 алгоритмінің шығыс биттері өзгеру ықтималдықтары таралуының интегралдық статистикалық көрсеткіштері

Көрсеткіш	**Мәндері**
Минимум	0.3367
Максимум	0.5943
Орташа мән	0.5061
Медиана	0.50
Мода	0.50
Дисперсия	0.002

Бақыланған ықтималдықтың ең төмен мәні 0.3367, ал ең жоғары мәні 0.5943 құрайды, яғни матрицаның жекелеген элементтері идеал 0.5 деңгейінен едәуір ауытқуы мүмкін. Алайда таралудың статистикалық сипаттамалары мұндай жағдайлардың бірлі-жарым екенін және жалпы көрініске әсер етпейтінін көрсетеді. Орташа мән 0.5061-ге тең, медиана – 0.50, мода да 0.50-ге тең, бұл таралудың теориялық мәнге қатысты симметриялыға жақын екенін көрсетеді. Шамамен 0.002 дисперсия мәндердің орташа шашырауын білдіреді және ықтималдықтардың көпшілігі 0.5 нүктесінің тар маңында жататынымен сәйкес келеді.

Жиынтық түрде бұл FLUX-128 алгоритмі үшін лавиндық ықтималдықтар орта есеппен оңтайлы деңгей маңында шоғырланғанын, ал байқалған ауытқулар жергілікті сипатта болып, айқын жүйелі асимметрия қалыптастырмайтынын растайды. Сандық сипаттамалар, лавиндық ауытқу диаграммалары түріндегі визуализация және ықтималдықтар матрицаларын талдау FLUX-128 алгоритмінің қатаң лавиндық критерийдің орындалуына жақын мінез-құлық көрсететінін өзара үйлесімді түрде көрсетеді.

4-кестеде берілген салыстырмалы талдау пайызбен өрнектелген лавиндық әсердің бірыңғай жинақталған метрикасы бойынша орындалды, мұнда 50%-ға жақын орташа мән жақсы диффузия деңгейінде күтілетін мінез-құлыққа сәйкес келеді. Blake-512, RIPEMD-160, SHA-512, MD5, Tiger, Grøstl-384 және Whirlpool стандартты хеш-функциялары бойынша деректер Upadhyay және авторлар зерттеуінен алынған, онда криптографиялық тұрғыдан берік хеш-функциялар жиыны үшін лавиндық әсер эксперименттік түрде зерттеліп, min, max және average жинақталған көрсеткіштері келтіріледі [12]. 4-кестедегі FLUX көрсеткіштері осы жұмыс аясында қалыптастырылған инверсия ықтималдықтары матрицасы бойынша SAC талдауының нәтижелерінен алынған.

Кесте 4 – FLUX-128 және криптографиялық хеш-функциялар үшін лавиндық әсердің салыстырмалы көрсеткіштері (Min, Max және Avg)

Хеш	**Min (%)**	**Max (%)**	**Avg (%)**
FLUX-128	33.67	59.43	50.61
Blake-512	43.36	58.79	51.07
Ripemd160	35.63	64.38	50
SHA-512	41.02	60.16	50.59
MD-5	32.81	65.63	49.22
Tiger	38.02	61.98	50
Grøstl-384	41.67	59.38	50.52
Whirlpool	41.8	57.03	49.42

4-кестеден көрініп тұрғандай, FLUX-128 үшін лавиндық әсердің орташа мәні 50.61% құрайды, яғни теориялық күтілетін 50% деңгейіне жақын және белгілі криптографиялық хеш-примитивтерге тән мәндермен салыстырмалы. Олармен салыстырғанда FLUX-128 диффузияның бәсекеге қабілетті интегралдық деңгейін көрсетеді.

4 Қорытынды

Жүргізілген зерттеу нәтижесінде ресурстары шектеулі есептеу жүйелерінде пайдалануға бағытталған жаңа FLUX-128 жеңілсалмақты хеш-функциясы құрастырылды. Алгоритм sponge тәсілі негізінде іске асырылған, бұл ретте оның архитектурасы классикалық шешімдермен салыстырғанда едәуір қайта өңделген. Атап айтқанда, G функциясын пайдалана отырып кіріс деректер ағынын қарапайымдатылған диффузиялық түрлендіру ұсынылды, ал шығыс мәнін қалыптастыру кезеңінде Davies–Meyer сұлбасы біріктірілді. Сонымен қатар жұмыста FLUX-128 алгоритмі биттік деңгейдегі диффузия сапасын және хабарламаларды сериялы өңдеу кезіндегі шығыс ағынының статистикалық мінез-құлқын көрсететін өзара толықтырушы екі бағыт бойынша эксперименттік бағаланды. SAC талдауының нәтижелері шығыс биттерінің инверсиялануының эмпирикалық ықтималдықтары орта есеппен теориялық күтілетін деңгей маңында шоғырланғанын көрсетті. Бұл таралудың сандық сипаттамаларымен де, матрицалар мен байланыстардың визуалды көріністерімен де расталады, мұнда кейінгі раундтарда ауытқулар құрылымының теңесуі байқалады.

Лавиндық әсердің жинақталған көрсеткіштерін белгілі криптографиялық хеш-примитивтермен салыстыру FLUX-128 диффузиясының орташа деңгейі бәсекеге қабілетті аймақта орналасқанын және эталондық 50% мәніне жақын екенін көрсетті.

Алғыс

Бұл зерттеу ҚР ҒК ҒЖБМ Ақпараттық және есептеу технологиялары институтында BR24993052 «Ресурстары шектеулі жүйелердегі ақпаратты қорғау үшін криптографиялық алгоритмдерді әзірлеу және зерттеу, олардың беріктігін бағалау» жобасы аясында орындалды.

Әдебиеттер

1. Turan, M.S., McKay, K.A., Chang, D., Bassham, L.E., Kang, J., Waller, N.D., Kelsey, J.M., Hong, D. *Status Report on the Final Round of the NIST Lightweight Cryptography Standardization Process*, NIST Interagency Report 8454, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. https://doi.org/10.6028/NIST.IR.8454.

2. Turan, M.S., McKay, K.A., Chang, D., Kang, J., Kelsey, J.M. *Ascon-Based Lightweight Cryptography Standards for Constrained Devices: Authenticated Encryption, Hash, and Extendable Output Functions*, NIST Special Publication 800-232, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2025. https://doi.org/10.6028/NIST.SP.800-232.

3. National Institute of Standards and Technology. *Advanced Encryption Standard (AES)*, Updated 9 May 2023, Federal Information Processing Standards Publication 197, U.S. Department of Commerce: Washington, DC, USA, 2001. https://doi.org/10.6028/NIST.FIPS.197-upd1.

4. Kerimbayeva, A., Iavich, M., Begimbayeva, Y., Gnatyuk, S., Tynymbayev, S., Temirbekova, Z., Ussatova, O. A Lightweight Variant of Falcon for Efficient Post-Quantum Digital Signature. *Information* 2025, *16*, 564. https://doi.org/10.3390/info16070564.

5. Kapalova, N., Algazy, K., Haumen, A., Sakan, K. Statistical analysis of the key scheduling of the new lightweight block cipher. *Int. J. Electr. Comput. Eng.* 2023, *13*, 6817–6826. https://doi.org/10.11591/ijece.v13i6.pp6817-6826.

6. Yan, L., Li, L., Guo, Y. DBST: A lightweight block cipher based on dynamic S-box. *Front. Comput. Sci.* 2023, *17*, 173805. https://doi.org/10.1007/s11704-022-1677-5.

7. Li, L., Liu, J., Guo, Y., Liu, B. A new S-box construction method meeting strict avalanche criterion. *J. Inf. Secur. Appl.* 2022, *66*, 103135. https://doi.org/10.1016/j.jisa.2022.103135.

8. Tolpa, S.H., Abdelhamed, M.A., Said, E.-S.S.A., Afi, M.Y.I. A novel chaos-based approach for constructing lightweight S-boxes with strong cryptographic properties. *Sci. Rep.* 2025, *15*, 34112. https://doi.org/10.1038/s41598-025-20019-4.

9. Madarro-Capó, E.J., Ramos Piñón, E.C., Sosa-Gómez, G., Rojas, O. Practical improvement in the implementation of two avalanche tests to measure statistical independence in stream ciphers. *Computation* 2024, *12*, 60. https://doi.org/10.3390/computation12030060.

10. Bellini, E., Huang, Y.J. *Randomness Testing of the NIST Light Weight Cipher Finalist Candidates*, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022. Available online: https://csrc.nist.gov/csrc/media/Presentations/2022/randomness-testing-of-the-nist-light-weight-cipher/images-media/session-6-huang-randomness-testing-of-lwc-cipher-finalists.pdf (accessed on 10 January 2026).

11. Khan, M., Johansen, D., Dagenborg, H. A comparative analysis of lightweight hash functions using AVR ATXMega128 and ChipWhisperer. *arXiv* 2025, https://doi.org/10.48550/arXiv.2508.07840.

12. Upadhyay, D., Gaikwad, N., Zaman, M., Sampalli, S. Investigating the avalanche effect of various cryptographically secure hash functions and hash-based applications. *IEEE Access* 2022, *10*, 112472–112486. https://doi.org/10.1109/ACCESS.2022.3215778.

УДК 004.056.5

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРОВ ЦИКЛИЧЕСКИХ СДВИГОВ НА ДИФФУЗИОННЫЕ СВОЙСТВА БЛОЧНОГО ШИФРА FEATHER-RX

Хаумен А.[1,2], Алгазы К.Т.[2]

e-mail: haumen.armanbek@gmail.com, kunbolat@mail.ru,

[1]Казахский Национальный Университет имени Аль-Фараби, Алматы, Казахстан

[2]Институт информационных и вычислительных технологий, Алматы, Казахстан

Аннотация. *В работе рассматривается легковесный блочный шифр FEATHER-RX и исследуется влияние параметров циклических сдвигов на его диффузионные свойства. Сравниваются различные варианты параметров вращения, анализируются лавинный эффект при локальном изменении открытого текста и чувствительность шифртекста к изменению секретного ключа. Дополнительно проводится предварительный анализ распространения входных разностей на сокращенном числе раундов и статистическая проверка выходных последовательностей. Полученные результаты позволяют сопоставить исследуемые конфигурации и обосновать выбор параметров вращения для дальнейшего исследования алгоритма. Работа не претендует на комплексный криптоанализ, а формирует экспериментальную основу для последующего расширенного анализа FEATHER-RX.*

Ключевые слова: *легковесная криптография, FEATHER-RX, сеть Фейстеля, циклический сдвиг, диффузия, лавинный эффект, дифференциальный анализ.*

1. ВВЕДЕНИЕ

Развитие Интернета вещей (Internet of Things, IoT), беспроводных сенсорных сетей и встраиваемых устройств повышает требования к защите данных в условиях ограниченных вычислительных ресурсов. Для таких платформ существенное значение имеют объем памяти, энергопотребление, задержка и сложность аппаратной или программной реализации. Это определило устойчивый интерес к легковесной криптографии и ее стандартизации для ресурсоограниченных устройств [1].

К числу известных легковесных блочных шифров относятся PRESENT, ориентированный на минимальные аппаратные затраты [2], а также семейства SIMON/SPECK, разработанные для гибкой реализации на ограниченных платформах [3]. Последующие конструкции, в частности SIMECK и GIFT, развивали идеи компактных раундовых преобразований и оптимизации аппаратной эффективности [4], [5]. Эти работы показывают, что выбор простых битовых операций и параметров раундовой функции является существенной частью проектирования легковесного шифра.

В легковесных конструкциях класса AND-RX широко применяются операции XOR, AND и циклические сдвиги. Исследования SIMON-подобных шифров показывают, что значения вращений влияют на распространение разностей и RX-характеристик, поэтому параметры сдвигов следует рассматривать как один из факторов, определяющих диффузионное и криптоаналитическое поведение раундовой функции [6]–[8].

В настоящей работе исследуется актуальная 36-раундовая версия легковесного блочного шифра FEATHER-RX. Алгоритм имеет 64-битный блок и 128/256-битный секретный ключ, построен на основе сбалансированной сети Фейстеля и использует комбинацию операций XOR, AND и циклических сдвигов. Одной из особенностей конструкции является применение изменяющихся параметров вращения, предназначенных для ускорения распространения изменений между битами состояния при сохранении простой структуры раундового преобразования.

Основное внимание уделяется экспериментальному выбору параметров циклических сдвигов и оценке их влияния на диффузионные свойства FEATHER-RX. Исследование имеет намеренно ограниченный характер и не рассматривается как комплексная оценка криптографической стойкости алгоритма. Его задача состоит в том, чтобы отдельно изучить поведение диффузии при выбранной конфигурации вращений и получить экспериментальные данные, которые впоследствии могут быть дополнены автоматизированным и теоретическим криптоанализом.

Для этого оценивается лавинный эффект при однобитном изменении открытого текста, а также чувствительность шифртекста к однобитному изменению секретного ключа. Дополнительно выполняется предварительный эксперимент по распространению структурированных входных разностей на сокращенном числе раундов. Статистические свойства выходных последовательностей проверяются с использованием NIST SP 800-22 [9]; при этом прохождение статистических тестов рассматривается только как характеристика распределения выходных данных и не интерпретируется как доказательство криптографической стойкости.

Основной вклад работы состоит в следующем: 1) предлагается экспериментальный подход к выбору параметров циклических сдвигов FEATHER-RX; 2) исследуется динамика диффузии при изменении одного бита открытого текста и секретного ключа; 3) выполняется предварительная оценка распространения входных разностей на сокращенном числе раундов, дополненная статистической проверкой выходных последовательностей. Полученные результаты предназначены для обоснования выбранной конфигурации FEATHER-RX и формирования основы для последующего расширенного исследования алгоритма.

2. FEATHER-RX И ВЫБОР ПАРАМЕТРОВ ЦИКЛИЧЕСКИХ СДВИГОВ

2.1. Краткое описание FEATHER-RX

FEATHER-RX представляет собой легковесный блочный шифр, построенный на основе сбалансированной сети Фейстеля. В рассматриваемой конфигурации размер блока составляет 64 бита, размер секретного ключа – 128/256 бит, а число раундов – 36. Входной блок разделяется на два 32-битных слова:

$$P = L_0 \parallel R_0, \tag{1}$$

где L_0 и R_0 обозначают левую и правую половины открытого текста. По набору используемых элементарных операций раундовая функция относится к AND-RX-подобным конструкциям: нелинейность формируется побитовым AND, а распространение зависимостей обеспечивается XOR и циклическими сдвигами. Подобный набор операций используется, в частности, в семействах SIMON и SIMECK [3], [4].

Раундовая функция FEATHER-RX определяется выражением

$$f_{RX}(R_i, i) = (ROTL_{32}(R_i, 1) \wedge ROTL_{32}(R_i, 8)) \oplus ROTL_{32}(R_i, 3) \oplus ROTL_{32}(R_i, t_i), \tag{2}$$

где $ROTL_{32}(x, s)$ обозначает циклический сдвиг 32-битного слова *x* влево на *s* позиций, а t_i – изменяемый параметр дополнительного вращения. Преобразование одного раунда имеет вид

$$\begin{aligned} L_{i+1} &= R_i, \\ R_{i+1} &= L_i \oplus f_{RX}(R_i, i) \oplus K_i, \end{aligned} \tag{3}$$

где K_i – 32-битный раундовый ключ. Фиксированные сдвиги на 1, 8 и 3 позиции сохраняются во всех раундах, тогда как дополнительное вращение изменяется циклически. В окончательной 36-раундовой конфигурации используется последовательность

$$R_{cycle} = (9,\ 19,\ 27), \quad t_i = R_{cycle}[i \bmod 3] \tag{4}$$

Таким образом, значения дополнительного вращения повторяются по раундам в порядке 9, 19, 27. В настоящей работе ключевое расписание подробно не рассматривается, поскольку основной предмет исследования связан с влиянием параметров вращения на диффузионные свойства раундового преобразования.

2.2. Выбор параметров вращения

Для AND-RX-конструкций выбор величин циклических сдвигов имеет криптографическое значение. Исследования SIMON-подобных шифров показывают, что конкретные значения вращений влияют как на распространение обычных разностей, так и на вероятность RX-характеристик [6], [8]. В работе Lu и соавторов RX-криптоанализ был непосредственно распространен на AND-RX-шифры SIMON и SIMECK; авторы отдельно отмечают влияние параметров вращения на распространение RX-характеристик [6]. Поэтому дополнительное вращение в FEATHER-RX выбиралось не как произвольный параметр реализации, а на основе экспериментального сравнения кандидатных конфигураций.

В качестве одного из исходных вариантов использовалась последовательность

$$R_{old} = (15,\ 21,\ 29) \qquad (5)$$

При исследовании вращательной зависимости на сокращенном числе раундов для этой конфигурации наблюдаемое смещение составляло приблизительно 0,00122. После анализа альтернативных комбинаций была выбрана последовательность

$$R_{final} = (9,\ 19,\ 27) \qquad (6)$$

для которой при тех же условиях значение наблюдаемого смещения снизилось приблизительно до 0,00011; наибольшее остаточное отклонение было зафиксировано для 37-го битового положения. Сравнение двух конфигураций приведено в таблице 1.

Таблица 1. Сравнение параметров динамического вращения FEATHER-RX

Набор вращений	Наблюдаемое смещение	Результат
(15, 21, 29)	≈ 0,00122	исходная конфигурация
(9, 19, 27)	≈ 0,00011	выбранная конфигурация

Полученные результаты показывают уменьшение наблюдаемой вращательной зависимости для комбинации (9, 19, 27) по сравнению с исходным вариантом. На основании этого результата, совместно с последующей оценкой диффузионных характеристик, данная комбинация была принята для 36-раундовой версии FEATHER-RX.

При этом приведенное сравнение не доказывает глобальную оптимальность набора (9, 19, 27) среди всех возможных комбинаций циклических сдвигов. Вывод ограничивается исследованными кандидатными конфигурациями и используемым экспериментальным критерием. Полный поиск оптимальных параметров и формальный анализ RX-характеристик выходят за рамки настоящей работы.

3. МЕТОДИКА ЭКСПЕРИМЕНТОВ

Экспериментальная оценка 36-раундовой версии FEATHER-RX проводилась по трем направлениям: исследование лавинного эффекта и чувствительности шифртекста к изменению секретного ключа, предварительный анализ распространения входных разностей на сокращенном числе раундов и статистическое тестирование выходных последовательностей. В качестве сравнительного алгоритма в основных экспериментах использовался SIMON64/128, имеющий тот же размер блока и ключа и использующий близкий набор элементарных битовых операций [3].

3.1. Диффузия при изменении открытого текста и ключа

Для оценки диффузии использовалось расстояние Хэмминга между состояниями шифра, полученными для исходного и модифицированного входа:

$$HD(C, C') = HW(C \oplus C') \quad (7)$$

В режиме изменения открытого текста для 1000 независимых пар (K, P) каждый из 64 битов P инвертировался отдельно, что давало 64 000 сравнений на одну контрольную точку. Промежуточные состояния фиксировались после 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 16, 20, 24, 28, 32 и 36 раундов. Для случайных независимых 64-битных состояний ожидается mean HD ≈ 32 и σ ≈ 4.

В режиме изменения ключа открытый текст сохранялся неизменным, а каждый из 128 битов секретного ключа инвертировался отдельно. Для 1000 исходных пар это соответствовало 128 000 сравнениям на одну контрольную точку. Полученный показатель рассматривается как чувствительность шифртекста к изменению ключевого материала, поскольку отражает совместное влияние ключевого расписания и раундового преобразования.

3.2. Проверка вращательной зависимости

Для проверки сохранения простых вращательных связей исследовались пары, в которых открытый текст и секретный ключ одновременно циклически сдвигались на одинаковую величину t ∈ {1, 3, 7, 9, 13, 19, 27, 31}. Для каждого значения t в 128-битном режиме использовалось 50 000 случайных пар. Особое внимание уделялось t = 9, 19 и 27, поскольку эти значения образуют динамический цикл вращения FEATHER-RX. На каждом выбранном раунде оценивалось расстояние Хэмминга между соответствующими вращательно связанными состояниями.

3.3. Двухэтапный анализ входных разностей

Дифференциальный эксперимент проводился в два этапа. На первом этапе рассматривались сокращенные версии с 2, 3, 4, 5, 6 и 8 раундами и 430 входных разностей, включавших все 64 однобитовые разности, 224 ориентированные на вращения двухбитовые разности, структурированные и случайные варианты. Для каждой разности и каждого режима ключа использовалось 100 000 пар открытых текстов.

На втором этапе 14 наиболее информативных кандидатов были повторно проверены на увеличенной выборке: для раундов 4, 5, 6 и 8 использовалось по 1 000 000 пар на каждую конфигурацию. Основным показателем служила максимальная наблюдаемая вероятность выходной разности

$$P_{\max} = \max_{\Delta C} \left[\frac{N(\Delta C)}{N} \right] \quad (8)$$

Такой двухэтапный подход позволяет сначала выполнить широкий поиск потенциально неблагоприятных входных разностей, а затем проверить устойчивость обнаруженных эффектов на существенно большей выборке. Значения, соответствующие единичным повторениям в миллионной выборке, интерпретируются как предел экспериментального разрешения, а не как точная оценка истинной максимальной дифференциальной вероятности.

Поэтому достижение значений вблизи указанного уровня интерпретируется как достижение экспериментального предела обнаружения, а не как определение истинной максимальной дифференциальной вероятности алгоритма. Представленный эксперимент является предварительной эмпирической оценкой и не заменяет полный поиск дифференциальных характеристик с использованием MILP, SAT или SMT-методов [8].

3.3. Статистическое тестирование выходных последовательностей

Статистические свойства шифртекстов дополнительно оценивались с использованием пакета NIST SP 800-22 [9]. Для тестирования формировались 64 независимые бинарные последовательности длиной 10^6 бит. Последовательности строились путем объединения шифртекстов, получаемых при последовательном

изменении входных блоков; для различных последовательностей использовался независимый ключевой материал.

Статистическая проверка проводилась при уровне значимости $\alpha = 0,01$.

Для каждого теста анализировались значения p-value и доля последовательностей, удовлетворяющих установленным NIST критериям. Такой подход позволяет выявить выраженные статистические отклонения, периодичность или неравномерность распределения выходных битов.

При этом прохождение тестов NIST SP 800-22 не рассматривается в настоящей работе как доказательство криптографической стойкости FEATHER-RX. Полученные результаты используются исключительно как дополнительная статистическая характеристика выходных последовательностей и рассматриваются совместно с анализом диффузии и распространения входных разностей.

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Экспериментальные результаты для 36-раундовой конфигурации FEATHER-RX показывают быстрое формирование диффузии как при изменении открытого текста, так и при изменении 128-битного секретного ключа. При инверсии одного бита открытого текста среднее расстояние Хэмминга возрастает от 2,50 бита после первого раунда до 31,57 после шестого и 31,98 после седьмого раунда. На полном 36-м раунде получено mean HD = 31,9872 при σ = 3,9856, что близко к модели двух независимых 64-битных состояний.

При однобитовом изменении ключа аналогичный статистический уровень достигается после 6–7 раундов: mean HD составляет 31,8735 на шестом и 32,0042 на седьмом раунде. На 36-м раунде mean HD = 31,9950 при σ = 3,9989. Результат характеризует совокупное распространение изменения ключевого материала через ключевое расписание и раундовое преобразование.

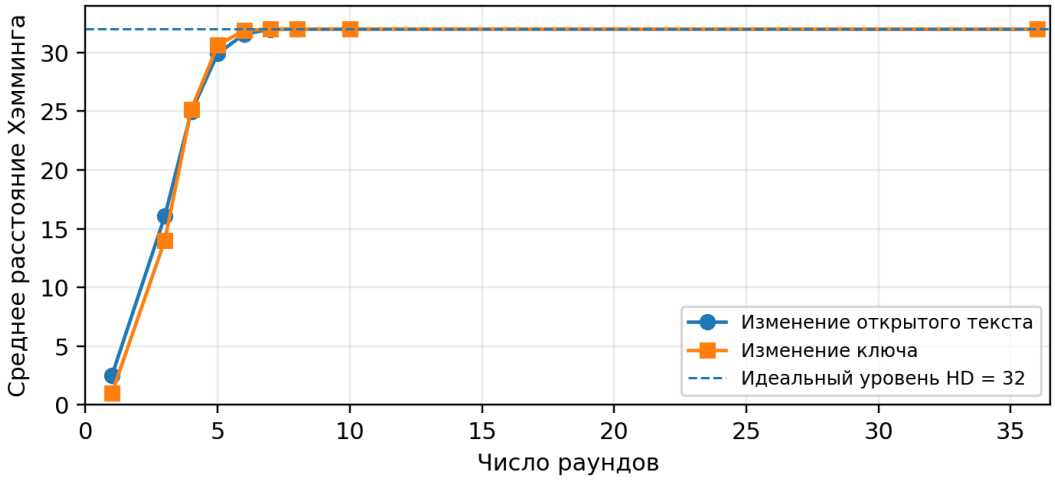


Рисунок 1. Формирование диффузии FEATHER-RX-36 при 128-битном ключе

Отдельный тест вращательной зависимости проводился для t ∈ {1, 3, 7, 9, 13, 19, 27, 31}. В 128-битном режиме на полном числе раундов mean HD по исследованным вращениям находился в диапазоне 31,9551–32,0269, а σ – приблизительно 3,988–4,016; случаев HD≤12 не зарегистрировано. В том числе для значений 9, 19 и 27, входящих в динамический цикл FEATHER-RX, устойчивой низковесной вращательной связи не обнаружено.

Дифференциальные эксперименты подтверждают быстрое ослабление доминирующих переходов на сокращенных версиях. Для 128-битного режима на четырех раундах максимальная наблюдаемая вероятность среди исследованных кандидатов составила $7,1\times10^{-4}$, тогда как на шести раундах в расширенной выборке она снизилась до 2×10^{-6}, а на восьми раундах зарегистрировано лишь одно повторение

на миллион пар (Таблица 2). Последняя точка соответствует пределу разрешения данного эксперимента и не должна интерпретироваться как точная оценка истинной максимальной дифференциальной вероятности.

Таблица 2. Выбранные результаты дифференциального эксперимента для 128-битного режима

Раунд	Этап	N	p_max	mean HD
4	Stage 1	100 000	$7{,}1\times10^{-4}$	22,167
6	Stage 2	1 000 000	2×10^{-6}	31,737
8	Stage 2	1 000 000	1×10^{-6}	31,998

В совокупности результаты указывают на быстрое распространение локальных изменений и ослабление исследованных вращательных и дифференциальных зависимостей с ростом числа раундов. При этом выводы имеют экспериментальный характер и не заменяют формальный поиск характеристик и теоретическую оценку стойкости.

Дополнительная статистическая оценка выходных последовательностей была проведена с использованием NIST SP 800-22 (NIST STS 2.1.2). Для каждого размера ключа были сформированы 64 независимые последовательности длиной 10^6 бит, что соответствует суммарному объему 128 млн бит для двух исследованных режимов. В 128-битном режиме минимальная доля успешно прошедших последовательностей составила 61/64, а в 256-битном режиме – 60/64, что не ниже установленного для данной выборки порога принятия. Для тестов Random Excursions также не было зарегистрировано результатов ниже соответствующих порогов. Минимальные итоговые значения вероятности составили 0,001399 и 0,000227 для 128- и 256-битных вариантов соответственно и оставались в области принятия. Таким образом, в исследованном объеме данных NIST SP 800-22 не выявил выраженных статистических отклонений в выходных последовательностях FEATHER-RX; при этом данный результат рассматривается только как дополнительная статистическая характеристика и не является доказательством криптографической стойкости алгоритма.

5. ЗАКЛЮЧЕНИЕ

В работе проведена экспериментальная оценка 36-раундовой конфигурации легковесного блочного шифра FEATHER-RX с акцентом на влияние параметров циклических сдвигов на диффузионные свойства алгоритма. По результатам предварительного сравнения в качестве динамического цикла вращения была выбрана последовательность (9,19,27), показавшая меньшую наблюдаемую вращательную зависимость по сравнению с исходной конфигурацией.

Эксперименты с однобитным изменением открытого текста и секретного ключа показали быстрое распространение локальных изменений по состоянию шифра. Для 128-битного режима среднее расстояние Хэмминга достигает уровня, близкого к 32 битам, приблизительно к 6–7 раундам и сохраняется около этого значения вплоть до полного 36-го раунда. Проверка вращательной зависимости не выявила устойчивых низковесных связей для исследованных величин сдвига, включая значения 9, 19 и 27, используемые в динамическом цикле.

Двухэтапный экспериментальный анализ входных разностей показал быстрое снижение максимальной наблюдаемой вероятности доминирующих выходных переходов с увеличением числа раундов. В расширенной выборке на восьми раундах было зарегистрировано только одно повторение одинаковой выходной разности на миллион пар, что соответствует пределу разрешения проведенного эксперимента и не рассматривается как точная оценка истинной максимальной дифференциальной вероятности.

Дополнительное тестирование выходных последовательностей с использованием NIST SP 800-22 не выявило статистических отклонений, выходящих за установленные критерии принятия. Для исследованных последовательностей показатели прохождения

тестов оставались в допустимой области, что подтверждает удовлетворительные статистические свойства выходных данных в рамках проведенного эксперимента.

Таким образом, полученные результаты подтверждают хорошие экспериментальные диффузионные и статистические характеристики выбранной конфигурации FEATHER-RX и быстрое ослабление исследованных вращательных и дифференциальных зависимостей при увеличении числа раундов. Вместе с тем проведенное исследование не является исчерпывающим доказательством криптографической стойкости. Дальнейшая работа должна включать автоматизированный поиск дифференциальных и линейных характеристик, анализ других классов криптоаналитических атак и более полную оценку запаса стойкости алгоритма.

Благодарность

Настоящее исследование выполнено в рамках проекта BR24993052 «Разработка и исследование криптографических алгоритмов для защиты информации в системах с ограниченными ресурсами и оценка их безопасности» в Институте информационных и вычислительных технологий Комитета науки Министерства науки и высшего образования Республики Казахстан.

Литература

[1] K. McKay, L. Bassham, M. Sönmez Turan, N. Mouha, “Report on Lightweight Cryptography,” NIST Interagency/Internal Report 8114, National Institute of Standards and Technology, 2017. doi: 10.6028/NIST.IR.8114.

[2] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. B. Robshaw, Y. Seurin, C. Vikkelsoe, “PRESENT: An Ultra-Lightweight Block Cipher,” in Cryptographic Hardware and Embedded Systems – CHES 2007, LNCS 4727, pp. 450–466, 2007. doi: 10.1007/978-3-540-74735-2_31.

[3] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, L. Wingers, “The SIMON and SPECK Lightweight Block Ciphers,” in Proceedings of the 52nd Annual Design Automation Conference (DAC), pp. 175:1–175:6, 2015. doi: 10.1145/2744769.2747946.

[4] G. Yang, B. Zhu, V. Suder, M. D. Aagaard, G. Gong, “The Simeck Family of Lightweight Block Ciphers,” in Cryptographic Hardware and Embedded Systems – CHES 2015, LNCS 9293, pp. 307–329, 2015. doi: 10.1007/978-3-662-48324-4_16.

[5] S. Banik, S. K. Pandey, T. Peyrin, Y. Sasaki, S. M. Sim, Y. Todo, “GIFT: A Small Present – Towards Reaching the Limit of Lightweight Encryption,” in Cryptographic Hardware and Embedded Systems – CHES 2017, LNCS 10529, pp. 321–345, 2017. doi: 10.1007/978-3-319-66787-4_16.

[6] J. Lu, Y. Liu, T. Ashur, B. Sun, C. Li, “Rotational-XOR Cryptanalysis of Simon-Like Block Ciphers,” in Information Security and Privacy – ACISP 2020, pp. 105–124, 2020. doi: 10.1007/978-3-030-55304-3_6.

[7] D. Khovratovich, I. Nikolić, “Rotational Cryptanalysis of ARX,” in Fast Software Encryption – FSE 2010, LNCS 6147, pp. 333–346, 2010. doi: 10.1007/978-3-642-13858-4_19.

[8] A. Biryukov, A. Roy, V. Velichkov, “Differential Analysis of Block Ciphers SIMON and SPECK,” in Fast Software Encryption – FSE 2014, pp. 546–570, 2015. doi: 10.1007/978-3-662-46706-0_28.

[9] L. Bassham, A. Rukhin, J. Soto, J. Nechvatal, M. Smid, S. Leigh, M. Levenson, M. Vangel, N. Heckert, D. Banks, “A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications,” NIST SP 800-22 Rev. 1a, 2010. doi: 10.6028/NIST.SP.800-22r1a.

ОЦЕНКА ЭФФЕКТИВНОСТИ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЦИФРОВОЙ ПОДПИСИ НА ОСНОВЕ ДЕРЕВА VERKLE

К.С. Сакан[1], С.Е. Нысанбаева[1], Маро[2] Е.А., Э.Б. Накипова[3]

*E-mail: kairat_sks@mail.ru

[1]Институт информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан

[2]Южный федеральный университет, Таганрог, Российскоя Федерация

[3]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан

Аннотация. *В статье исследуется экспериментальная оценка эффективности предложенной цифровой подписи на основе дерева Verkle с использованием китайской теоремы об остатках (КТО). Разработано программное обеспечение для алгоритмов генерации ключей, формирования подписи и ее проверки. В предлагаемой схеме дерево Verkle используется для компактного представления обязательств, а китайская теорема об остатках – для оптимизации модульных вычислений и повышения вычислительной эффективности операций формирования и проверки подписи. Проведен анализ временных характеристик алгоритмов и получены показатели вычислительной сложности. Для обеспечения статистической достоверности экспериментальные результаты получены на фиксированной вычислительной платформе при многократном выполнении тестов. Выполнен сравнительный анализ разработанного алгоритма цифровой подписи и классической схемы цифровой подписи на основе полиномиальных обязательств Kate–Zaverucha–Goldberg (KZG) по основным параметрам и времени выполнения. Полученные результаты показывают перспективность использования дерева Verkle в сочетании с китайской теоремой об остатках для построения цифровых подписей.*

Ключевые слова: *дерево Verkle, векторное обязательство, полиномиальное обязательство, китайская теорема об остатках, цифровая подпись, аутентификация, верификация.*

Введение

Современные схемы цифровой подписи, такие как Crystals-Dilithium, Falcon и SPHINCS+, демонстрируют высокий уровень криптографической стойкости. Однако для этих алгоритмов характерны относительно большие размеры ключей и подписей, а также значительные вычислительные затраты. Поэтому актуальным направлением является поиск альтернативных структурных решений, обеспечивающих уменьшение объема передаваемых данных и оптимизацию вычислительных операций.

Одним из перспективных направлений является структура Verkle, основанная на векторных обязательствах. Она обеспечивает логарифмическую сложность доказательства принадлежности элемента и компактность криптографического доказательства. Применение китайской теоремы об остатках предоставляет дополнительные возможности оптимизации вычислений, позволяя декомпозировать модульные операции и повысить эффективность арифметических вычислений.

Основной целью работы является разработка программного обеспечения для цифровой подписи на основе дерева Verkle с применением КТО, экспериментальная оценка ее эффективности и сравнительный анализ с современными схемами цифровой подписи.

Научная новизна заключается в разработке программного обеспечения и проведении экспериментального исследования цифровой подписи на основе дерева Verkle с использованием КТО. В отличие от исследований, где дерево Verkle и методы оптимизации модульной арифметики рассматриваются преимущественно раздельно,

здесь предлагается их объединение в рамках единой алгоритмической модели цифровой подписи.

Практическая значимость заключается в создании программного средства, позволяющего оценивать эффективность вычислений, ориентированных на КТО, в структурах на основе дерева Verkle. Разработанное ПО может использоваться для дальнейших исследований в области компактных криптографических протоколов и систем, требующих оптимизации модульной арифметики.

Выполнение операций с полиномами высокой степени над полем Галуа, особенно поиск мультипликативно обратного элемента, является вычислительно сложным процессом. Кроме того, проверка примитивности полиномов также увеличивает вычислительные затраты. Для повышения эффективности этих операций использованы оптимизированные подходы.

Обзор литературы

Разработка и анализ современных схем цифровой подписи активно развиваются в рамках перехода к постквантовой криптографии. Одним из значимых направлений являются решеточные схемы подписи, отобранные в рамках конкурса Национального института стандартов и технологий США (NIST) по стандартизации постквантовых алгоритмов. В частности, алгоритм Crystals-Dilithium [1] основан на сложности задач Module-LWE и Module-SIS. В работе приведены формальные определения алгоритмов генерации ключей, формирования и проверки подписи, обоснование параметров безопасности и практические аспекты реализации.

Работа M. Iavich и соавторов [2] рассматривает применение дерева Verkle в направлении постквантовой цифровой подписи. Авторы анализируют его особенности по сравнению с классическим деревом Меркла и отмечают преимущества уменьшения размера доказательства принадлежности за счет векторных обязательств и высокой степени ветвления. Показано, что структуры Verkle позволяют существенно уменьшить глубину дерева и объем подтверждающей информации при проверке.

В обзоре Khalfin [3] рассматриваются основные направления применения КТО в криптографии, включая ускорение вычислений в системах типа RSA, построение секретных систем и модульную декомпозицию сложных вычислительных операций. Применение КТО позволяет заменить вычисления по большому модулю набором параллельных операций по взаимно простым модулям с последующим восстановлением результата. При этом теорема рассматривается как инструмент оптимизации, а не как самостоятельная криптографическая предпосылка безопасности.

В работе Grossschädl [4] аппаратная реализация RSA-чипа с оптимизацией на основе КТО позволяет примерно в 3,5 раза ускорить операцию расшифрования за счет разделения модульного возведения в степень на две параллельные подзадачи. Это показывает практическую применимость КТО для повышения производительности криптографических вычислений.

Экспериментальные исследования оптимизации RSA показывают, что версия RSA с оптимизацией на основе КТО выполняет возведение в степень по модулю n примерно в 3–4 раза быстрее классических операций благодаря разбиению вычислений на параллельные подзадачи по меньшим модулям и последующему восстановлению результата [5].

В статье Demir и соавторов [6] анализируются эффективность и возможности промышленного внедрения алгоритмов постквантовой криптографии. Исследуются временные характеристики генерации ключей, формирования подписи и ее проверки, а также влияние алгоритмов на платформы с ограниченными ресурсами. Эти результаты подчеркивают необходимость экспериментальной оценки новых схем в конкретных вычислительных условиях.

Рассмотренные исследования охватывают разработку и стандартизацию постквантовых схем цифровой подписи, исследование структур обязательств либо

оптимизацию арифметики с помощью КТО. Однако комплексных исследований, объединяющих структуру дерева Verkle и применение КТО в рамках экспериментальной реализации цифровой подписи, сравнительно мало. Это определяет актуальность настоящей работы.

Основной вклад работы

- Интеграция полиномиальных обязательств на основе КТО в структуру дерева Verkle для эффективного представления и проверки принадлежности данных.
- Разработка полной алгоритмической модели цифровой подписи, включающей KeyGen, Commit, Open, Verify и Update, а также программная реализация предлагаемой схемы (VCNPSS v1.0) с использованием библиотек NTL и GMP.
- Экспериментальная оценка вычислительной эффективности в зависимости от параметров безопасности n и размера структуры k.
- Сравнительный анализ с классической схемой полиномиальных обязательств KZG по вычислительной сложности, структуре доказательств и особенностям реализации.

Материалы и методы исследования

Дерево Verkle

Основным преимуществом дерева Меркла является логарифмическая сложность доказательства принадлежности элемента. Однако с увеличением объема данных глубина дерева возрастает, что приводит к увеличению размера доказательства и числа хэш-операций при проверке.

В качестве альтернативы используется дерево Verkle, основанное на векторных обязательствах. В отличие от бинарной иерархии, оно допускает высокую степень ветвления и использует криптографические векторные обязательства для объединения нескольких значений в одном узле (рисунок 1). Это позволяет существенно уменьшить высоту дерева и более эффективно доказывать принадлежность элемента.

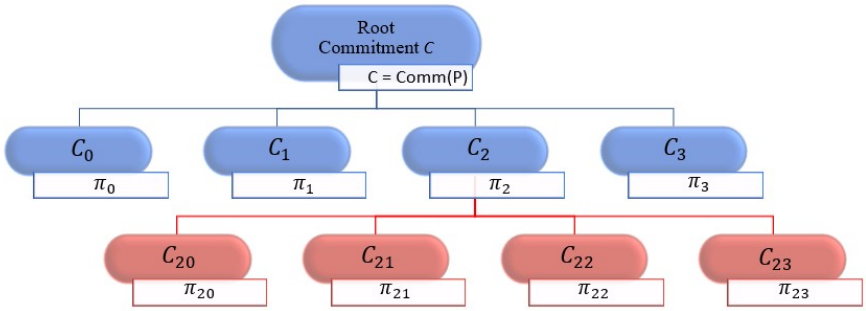


Рисунок 1 - Структурная схема дерева Verkle.

Использование дерева Verkle в схеме цифровой подписи уменьшает объем информации, требующей подтверждения, и сокращает число вычислительных операций при проверке. В сочетании с оптимизацией модульных вычислений на основе КТО эта структура формирует основу криптографической конструкции, адаптированной к обработке больших объемов данных.

Китайская теорема об остатках

Китайская теорема об остатках является одним из основных инструментов теории чисел, используемых для решения систем линейных сравнений по различным модулям. Если модули $m_1, m_2, \dots, m_n$ взаимно просты, то для любого набора целых чисел $a_1, a_2, \dots, a_n$ существует единственное число x по модулю $M = \prod_{i=1}^{n} m_i$, удовлетворяющее $x \equiv a_i (mod\ m_i),\ \ i = 1, 2, \dots, n$ [7]. Практическая значимость

теоремы заключается в возможности разложения вычислений по большому модулю на операции по меньшим взаимно простым модулям с последующим восстановлением общего результата. Такой подход снижает вычислительные затраты при работе с числами большой разрядности и применяется в криптографических алгоритмах, цифровой обработке сигналов и системах кодирования информации [8].

Процедура построения решения включает определение общего модуля $M = \prod_{i=1}^{n} m_i$, вычисление частичных модулей $M_i = \frac{M}{m_i}$ и нахождение мультипликативно обратных элементов yi, удовлетворяющих $M_i \cdot y_i \equiv 1(mod m_i)$. Решение формируется как линейная комбинация $x = \sum_{i=1}^{n} a_i M_i y_i$. В структуре дерева Verkle КТО позволяет представить вычисления по большому модулю M как набор независимых операций по взаимно простым модулям с последующим восстановлением результата, декомпозируя арифметические операции формирования и проверки векторных обязательств.

Результаты и их обсуждение

В ходе исследования разработана программная модель цифровой подписи, основанная на полиномиальных обязательствах в структуре дерева Verkle с использованием КТО. Алгоритмическая схема включает процедуры генерации ключей (KeyGen), формирования обязательств (Com), раскрытия доказательства (Open), проверки обязательства (Ver), а также обновления обязательств и доказательств (Update, ProofUpdate).

Исходные параметры

Для корректной программной реализации необходимо определить значение параметра безопасности *n*, число *k* узлов дерева Verkle, множество попарно взаимно простых неприводимых полиномов $p_i(x)$ заданной степени и набор ключей $g_i(x)$, сгенерированных генератором псевдослучайных последовательностей.

1) *n* – параметр безопасности, определяющий степень полинома или его длину в двоичном представлении.

2) $k \in N$ количество узлов в дереве Verkle.

3) С использованием генератора G формируется необходимое количество секретных ключей gi(x), образующих $G(x) = (g_1(x), g_2(x), ..., g_k(x))$, где $def\,(g_i(x)) = n$ и $g_1(x) \neq g_2(x) \neq ... \neq g_k(x)$, $i = 1..k$.

4) Формируется множество P неприводимых полиномов степени *n + 1*. Из него случайным образом выбираются $p_i(x)$, $p_1(x) \neq p_2(x) \neq ... \neq p_k(x)$, $i = 1..k$. Они рассматриваются как рабочие основания непозиционной полиномиальной системы счисления.

5) Для хэширования используется криптографически стойкая функция $H : \{0, 1\}^* \to \{0, 1\}^n$. В работе использована ранее разработанная программная реализация LWH-512. Выбор LWH-512 обусловлен стойкостью к атакам на коллизии, прообраз и второй прообраз, а также лавинным эффектом и устойчивыми статистическими характеристиками [9]. Использование 512-битного выходного значения при формировании доказательств πi(x) и вычислении полиномиальных обязательств обеспечивает необходимый уровень криптографической стойкости.

6) Для вычисления мультипликативно обратного полинома $P_i^{-1}(x)$ по модулю $p_i(x)$ используется расширенный алгоритм Евклида.

Алгоритм программной реализации

1) Этап генерации ключей – $KeyGen\,(1^n,\ k)$.

С использованием генератора *G* создаются ключи $g_i(x)$, $i = 1..k$, образующие полный ключ $G(x) = (g_1(x), g_2(x), ..., g_k(x))$.

1.1) Из множества $\mathbb{P}$ выбираются $p_i(x)$, $i = 1..k$.

2) Этап формирования обязательств и доказательств – $Com_{pp}\left(m_1(x),\dots,m_k(x)\right)$.

2.1) Для каждого $m_i(x)$ вычисляется $h_i(x)=H\left(m_i(x)\right)$.

2.2) Вычисляются $\pi_i(x)=h_i(x)\oplus g_i(x), \quad i=\overline{1,k}$, при этом $\deg\left(g_i(x)\right)=\deg\left(h_i(x)\right)=n$.

2.3) Для вычисления *C(x)* по КТО:

1. Вычисляются $P(x)=\prod_{i=1}^{k}p_i(x)$.

2. Вычисляются $P_i(x)=\frac{P(x)}{p_i(x)}$, $i=\overline{1,k}$,

3. С помощью алгоритма Эвклида находятся $P_i^{-1}(x)$, удовлетворяющие $P_i(x)\bullet P_i^{-1}(x)\equiv 1\ (mod\ p_i(x))$.

4. Обязательство вычисляется как $C(x)=\sum_{i=1}^{k}\pi_i(x)\bullet P_i(x)\bullet P_i^{-1}(x)\ mod\ P(x)$ и публикуется для подтверждения данных.

3) Этап раскрытия – $Open\left(m_i(x),\pi_i(x),\ i\right)$.

Публикуется доказательство того, что сообщение $m_i(x)$ имеет обязательство $\pi_i(x)$, входящее в *C(x)*, без раскрытия других значений $\pi_j(x)$, *i* ≠ *j*.

4) Этап верификации – $Ver\left(C(x),p_i(x),g_i(x),i\right)$.

4.1) Вычисляется $\pi_i(x)=C(x)\ mod\ p_i(x)$.

4.2) Вычисляются $y_i(x)=H\left(m_i(x)\right)\oplus g_i(x),\ i=\overline{1,k}$.

4.3) Если $\pi_i(x)=y_i(x)$, подтверждается корректность обязательства и неизменность сообщения.

5) Этап обновления – $Update\left(C(x),m_i(x),m_i'(x),g_i'(x),i\right)$.

При замене *i*-го сообщения обязательство пересчитывается как $C'(x)=\sum_{i=1}^{k}\pi_{\mathbf{i}}'(x)\bullet P_{\mathbf{i}}'(x)\bullet P_{\mathbf{i}}'^{-1}(x)\ mod\ P(x)$, $i=\overline{1,k}$. Повторное вычисление *P(x)* не требуется.

Предложенная последовательность операций отражает логическую и вычислительную структуру модели и обеспечивает корректность формирования и проверки обязательств. Описание процедур позволяет адаптировать программное обеспечение и экспериментально оценивать его эффективность.

Программное обеспечение алгоритма

Для практической реализации предложенной теоретической модели возникла необходимость разработки специализированного программного обеспечения, обеспечивающего функционирование алгоритмов формирования и проверки обязательств. В связи с этим в соответствии с описанной схемой был разработан программный комплекс VCNPSS v1.0, реализующий основные криптографические процедуры.

Программный комплекс VCNPSS v1.0 имеет модульную архитектуру и включает следующие основные компоненты:

- модуль генерации исходных параметров;
- модуль хэширования;
- модуль криптографических вычислений;
- модуль пользовательского интерфейса.

Разработанный программный комплекс VCNPSS v1.0 реализован на языке программирования C++ с использованием фреймворка Qt версии 5.15.18, библиотеки GMP (GNU Multiple Precision Arithmetic Library) версии 6.3.0, а также математической библиотеки NTL (Number Theory Library) версии 11.6.0.

Ниже приведены примеры пользовательского интерфейса, иллюстрирующие основные этапы работы программного обеспечения VCNPSS v1.0 для случая, когда степень неприводимых полиномов составляет 512 (рисунки 2–5).

В программном обеспечении используется библиотека NTL (Number Theory Library) версии 11.6.0, предназначенная для выполнения вычислений в области теории чисел и алгебры. Библиотека предоставляет широкие возможности для работы с большими целыми числами и полиномами над конечными полями, что позволяет реализовывать операции, необходимые для арифметики GF(2) и применения китайской теоремы об остатках.

Для повышения производительности библиотека NTL была собрана с поддержкой библиотеки GMP (GNU Multiple Precision Arithmetic Library), обеспечивающей арифметику произвольной точности. Использование данной связки позволяет ускорить операции модульных вычислений и обеспечить корректность работы с полиномами высокой степени.

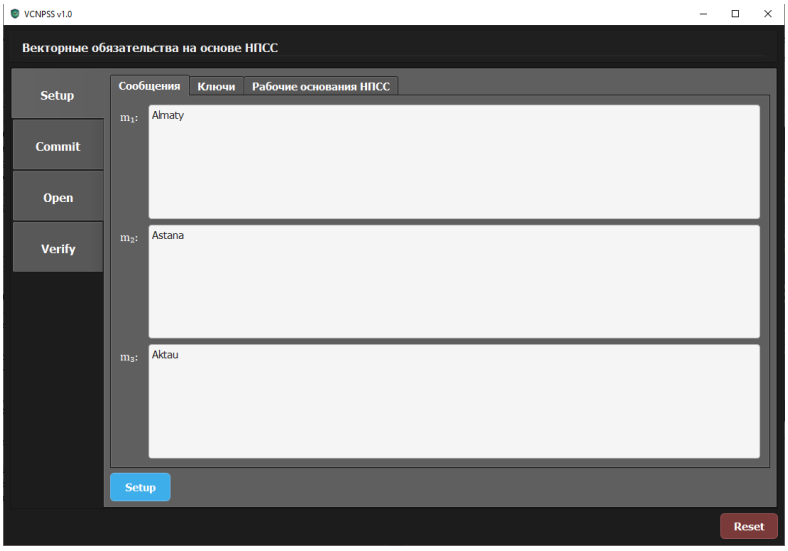


Рисунок 2 - Главное окно программного обеспечения VCNPSS v1.0 (этап Setup).

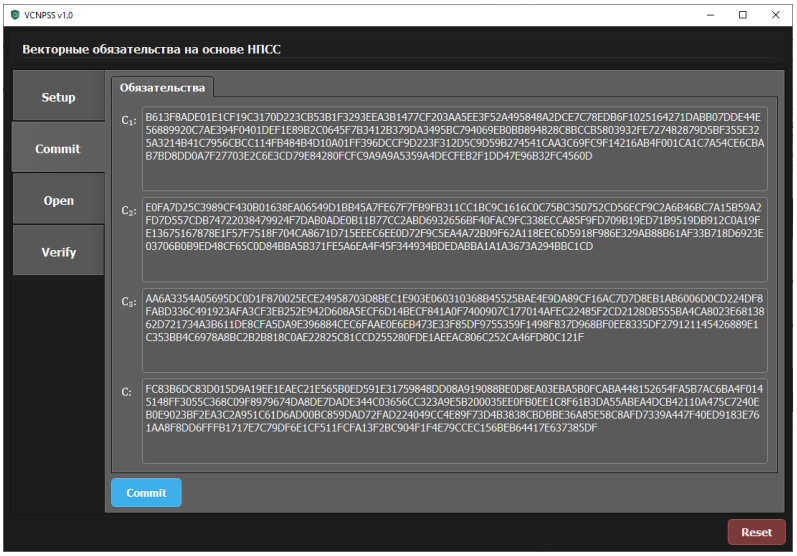


Рисунок 3 - Отображение сгенерированных криптографических обязательств (этап Commit).

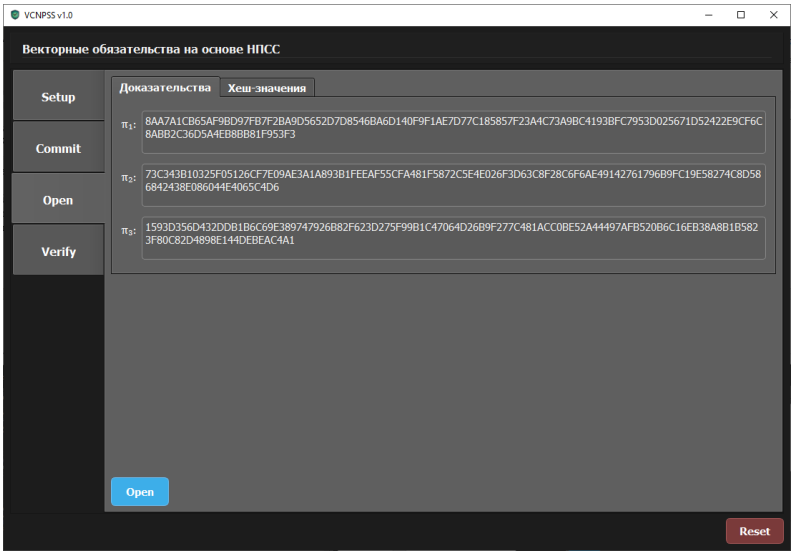


Рисунок 4 - Отображение криптографических доказательств принадлежности сообщений обязательствам (этап Open).

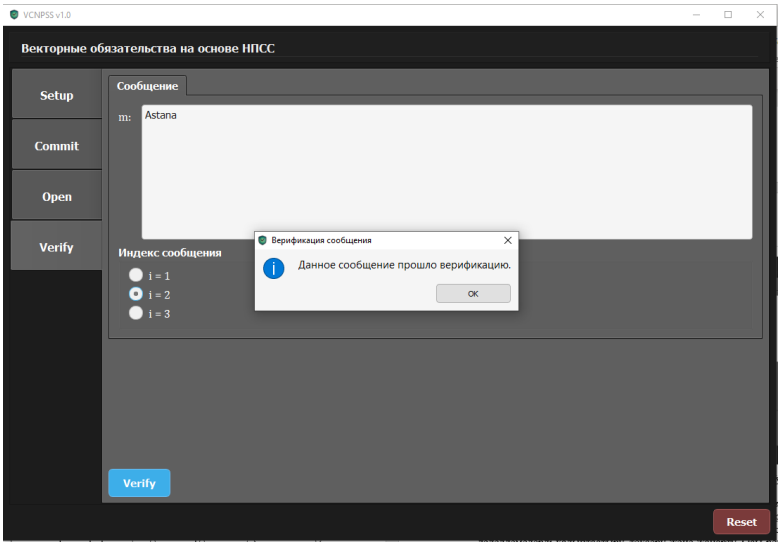


Рисунок 5 - Успешно верифицированное сообщение (этап Verify).

Библиотека NTL предоставляет возможности для работы с большими целыми числами и полиномами над конечными полями, что позволяет выполнять операции, необходимые для арифметики *GF(2)* и применения КТО. Для повышения производительности NTL собрана с поддержкой GMP, обеспечивающей арифметику произвольной точности.

Теоретическая оценка вычислительной эффективности алгоритма

Оценка вычислительной эффективности проведена на основе анализа сложности генерации ключей, формирования обязательств и доказательств, проверки и обновления. Особое внимание уделено операциям полиномиальной арифметики и реконструкции обязательства с использованием КТО [10].

1. Формирование доказательств. Для каждого сообщения m_i, $i = 1, \dots, k$ выполняются вычисление $H(m_i)$ и XOR с g_i. При фиксированной длине хэш-кода $n = 512$ сложность одного доказательства составляет $O(k)$, а для всех k узлов – $O(k \cdot n)$.

2. Вычисление обязательства. Наиболее ресурсоемким является вычисление $C = \left(\sum_{i=1}^{k} \pi_i P_i P_i^{-1}\right) \bmod P$. Вычисление произведения модулей требует $O(k \cdot n^2)$

операций, а поиск обратных элементов с помощью алгоритма Евклида – $O(n^2)$. Выполняются также модульные сложения и приведение по модулю. Итоговая асимптотическая сложность формирования обязательства оценивается как $O(kn^2)$.

3. Верификация. Процедура включает вычисление остатка, хэширование, XOR и сравнение значений. Сложность проверки одного сообщения составляет $O(n^2)$.

4. Использование КТО. Применение КТО позволяет заменить вычисления по модулю P набором независимых операций по модулям pi, обеспечивая декомпозицию задачи и возможность параллельной реализации. Поэтому практическое время выполнения может быть меньше, чем при прямых вычислениях по одному модулю высокой степени.

Сравнительный анализ с KZG

В таблице 1 приведен сравнительный анализ схемы полиномиальных обязательств на основе КТО и классической схемы KZG [11, 12]. Рассматриваются размер доказательства, вычислительная сложность формирования и проверки, основные операции, особенности реализации и постквантовая стойкость. Выбор KZG связан с тем, что предлагаемая схема также основана на полиномиальных обязательствах.

Таблица 1 – Сравнительный анализ с полиномиальной схемой KZG

Параметры	**Предлагаемая схема**	**KZG**
Основные вычисления	Полиномиальная арифметика, реконструкция по CRT	Эллиптические кривые и pairing вычисление
Криптографическая конструкция	Полиномиальное обязательство	Полиномиальное обязательство
Размер доказательства	Один полином πi степени n	Один элемент группы G1
Формирование обязательства	Работа с полиномами — *O(k·n²)*	Скалярное умножение — O(d)
Верификация	Работа с полиномами — *O(n²)*	Скалярное умножение — O(log d)
Рост производительности	Линейный по k	Зависит от степени полинома d
Основные сложные операции	Умножение полиномов, алгоритм Евклида	Scalar multiplication, bilinear pairing
Сложность реализации	Средняя	Высокая
Аппаратная зависимость	Очень низкая (обычная арифметика)	Высокая (эллиптические кривые и pairing)
Постквантовая стойкость	Да	Нет

Сравнительный анализ показывает, что вычислительная модель предлагаемой схемы основана главным образом на операциях модульной арифметики и реконструкции с использованием КТО, тогда как KZG использует ресурсоемкие операции на эллиптических кривых и билинейные отображения. В предлагаемой схеме вычислительная сложность определяется степенью полиномов и числом узлов k. Схема не использует эллиптические кривые и билинейные отображения; в рамках принятой модели она рассматривается как постквантово ориентированное решение.

Экспериментальная оценка эффективности

Экспериментальная оценка выполнена для процедуры формирования обязательства C, являющейся наиболее ресурсоемким этапом алгоритма. Измерения проводились для различных значений параметра безопасности n и числа узлов k с использованием КТО.

Программное обеспечение использовало GMP для поддержки арифметики произвольной точности. Основную долю вычислительной сложности составляют операции полиномиальной арифметики и реконструкция обязательства по КТО.

Полученные временные показатели отражают прежде всего алгоритмические свойства предлагаемой схемы.

При фиксированном значении *n = 512* увеличение числа узлов k приводит к линейному росту времени формирования обязательства. Это объясняется тем, что обязательство вычисляется как сумма *k* слагаемых вида $\pi_i P_i P_i^{-1}$. Каждый дополнительный узел требует набора операций умножения полиномов и последующего приведения по модулю $P(x)$ (рисунок 6).

Поскольку степень полиномов определяется параметром *n*, сложность каждой отдельной операции умножения и приведения по модулю остается постоянной по порядку величины. Рост вычислительной нагрузки определяется главным образом увеличением числа компонентов k, а зависимость времени выполнения от числа узлов имеет линейный характер.

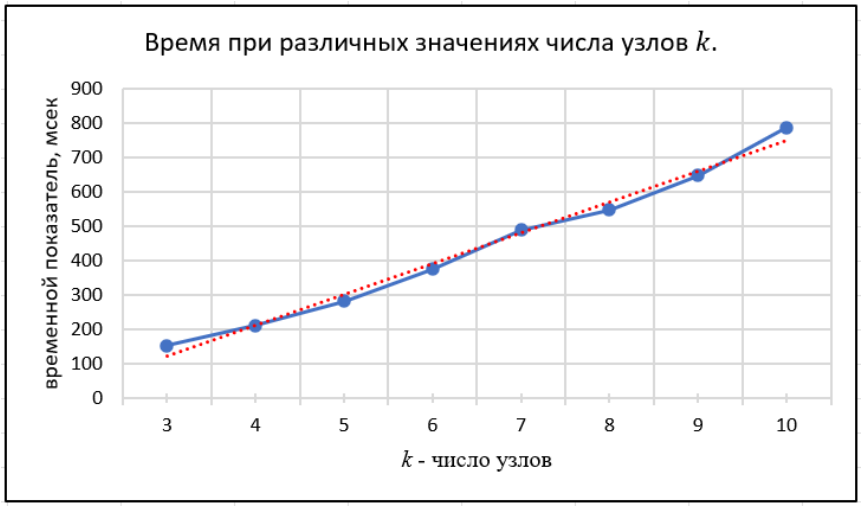


Рисунок 6 - Динамика времени при различных значениях k для n = 512.

Проведенные экспериментальные измерения подтверждают соответствие практических результатов теоретической оценке сложности *O(n²k)*. Это свидетельствует о возможности масштабирования алгоритма по параметру *k* при заданном уровне криптографической стойкости.

При фиксированном *k = 3* увеличение параметра безопасности n приводит к нелинейному росту времени формирования обязательства. Это объясняется ростом степени используемых полиномов и разрядности коэффициентов, что непосредственно влияет на умножение полиномов, деление по модулю и вычисление остатков (рисунок 7).

Нахождение элементов P_i^{-1} с использованием расширенного алгоритма Евклида также существенно влияет на производительность. С увеличением степени полиномов возрастает число итераций и объем промежуточных вычислений, что приводит к квадратичному росту вычислительных затрат. Увеличивается и количество операций реконструкции обязательства по КТО.

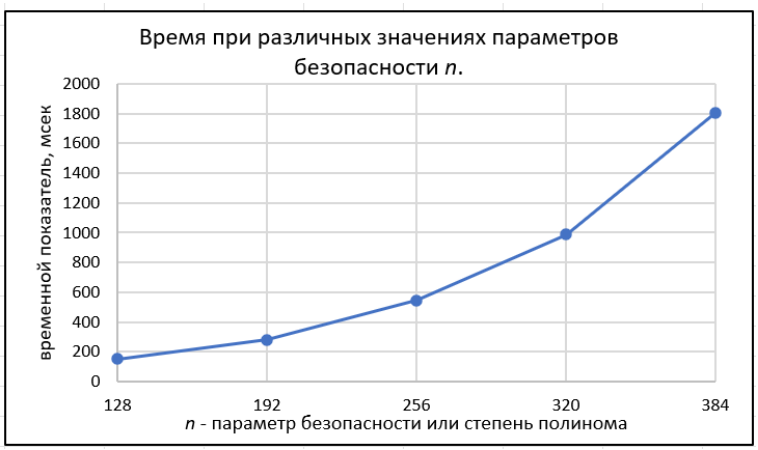


Рисунок 7 - Динамика времени при фиксированном $k = 3$ и различных значениях n.

Таким образом, увеличение параметра безопасности комплексно влияет на основные этапы формирования обязательства и объясняет наблюдаемую нелинейную динамику времени выполнения. Полученные результаты при фиксированном числе узлов согласуются с теоретической оценкой сложности порядка $O(n^2)$, а график характеризуется параболической формой.

В работе в качестве хэш-функции использована схема LWH. Алгоритм основан на криптографической губчатой (sponge) конструкции, позволяющей гибко изменять длину выходного значения: хэш-модель может быть расширена до 512 бит. Предлагаемая схема цифровой подписи не привязана к конкретной хэш-функции и допускает использование любой криптографической хэш-функции, устойчивой к коллизиям. Применение стандартизованных алгоритмов SHA-3 и SHAKE соответствует предложенной модели и может повысить уровень безопасности.

Заключение

В работе предложена схема формирования и проверки полиномиальных обязательств, основанная на применении полиномиальных структур и китайской теоремы об остатках в дереве Verkle. Разработанная модель позволяет формировать обязательства и доказательства принадлежности сообщений без раскрытия остальных элементов структуры. Применение КТО обеспечивает декомпозицию вычислений и возможность реконструкции полиномиального обязательства, что способствует упрощению арифметических операций и повышению эффективности обработки данных.

В ходе исследования реализована программная модель алгоритма с использованием математической библиотеки NTL и библиотеки высокоточной арифметики GMP. Анализ вычислительной эффективности показал зависимость времени формирования обязательства от параметра безопасности n и количества узлов k. При фиксированном n увеличение k приводит к линейному росту вычислительных затрат, тогда как увеличение n вызывает нелинейный рост времени выполнения, связанный с повышением степени полиномов и усложнением операций модульной арифметики.

Сравнительный анализ с классической схемой полиномиальных обязательств KZG показал, что предлагаемая схема использует полиномиальную арифметику и реконструкцию по КТО вместо операций с эллиптическими кривыми и билинейными отображениями. Алгоритм построен без групповых криптографических операций; в рамках принятой модели это позволяет рассматривать его как постквантово ориентированное решение.

Полученные результаты показывают перспективность применения полиномиальных обязательств на основе КТО для построения механизмов аутентификации, адаптированных к обработке больших объемов данных. Разработанная

схема может служить основой дальнейших исследований в области постквантовых цифровых подписей и других криптографических протоколов, использующих векторные обязательства.

Для валидации предложенных методов разработана специальная компьютерная программа. Программное средство обеспечивает реализацию модели, основанной на неприводимых полиномах степени $n = 512$.

Благодарность

Исследовательская работа выполнена в Институте информационных и вычислительных технологий МНВО РК в рамках проекта AP23488112 «Разработка и исследование схемы квантово-стойкой цифровой подписи на основе дерева Verkle».

Список литературы

1. Ducas, R., Kiltz, E., Lepoint, T., Lyubashevsky, V., Schwabe, P., Seilor, G. & Stehle, D. (2021). CRYSTALS-Dilithium: Algorithm specification (Version 3.1).

2. Iavich, M., Gagnidze, A. & Iashvili, G. (2023). Verkle tree-based post-quantum digital signature: Algorithms and applications. CEUR Workshop Proceedings, 3350, 1–12.

3. Khalfin, M. (2021). Chinese Remainder Theorem and Cryptography.

4. Grossschädl, J. (2000). The Chinese remainder theorem and its application in a high-speed RSA crypto chip. Proceedings of the 16th Annual Computer Security Applications Conference, 384–393. DOI: 10.1109/ACSAC.2000.898893.

5. Abid, R., Iwendi, C., Javed, A.R. et al. (2023). An optimised homomorphic CRT-RSA algorithm for secure and efficient communication. Personal and Ubiquitous Computing, 27, 1405–1418. DOI: 10.1007/s00779-021-01607-3.

6. Demir, E.D., Bilgin, B. & Onbasli, M.C. (2025). Performance analysis and industry deployment of post-quantum cryptography algorithms. arXiv:2503.12952. DOI: 10.48550/arXiv.2503.12952.

7. Algazy, K., Sakan, K. & Sawicki, D. (2025). A new digital signature scheme based on the Verkle tree using the Chinese remainder theorem. Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2025, 14009, 1–7. DOI: 10.1117/12.3099534.

8. Algazy, K., Sakan, K., Kapalova, N.A. & Alimzhan, Y.Zh. (2025). A new approach to constructing vector commitments for the Verkle tree. Vestnik KazUTB, 3, 104–116. DOI: 10.58805/kazutb.v.3.28-928.

9. Sakan, K., Algazy, K., Kapalova, N. & Varennikov, A. (2025). Lightweight hash function design for the Internet of Things: Structure and SAT-based cryptanalysis. Algorithms, 18(9), 550. DOI: 10.3390/a18090550.

10. Omondi, A. & Premkumar, B. (2007). Residue Number Systems: Theory and Implementation. Imperial College Press. DOI: 10.1142/p523.

11. Kate, A., Zaverucha, G. & Goldberg, I. (2010). Constant-size commitments to polynomials and their applications. ASIACRYPT 2010, 6477, 177–194. DOI: 10.1007/978-3-642-17373-8_11.

12. Gabizon, A., Williamson, Z. & Ciobotaru, O. (2019). PLONK: Permutations over Lagrange-bases for Oecumenical Noninteractive arguments of Knowledge. IACR Cryptology ePrint Archive, 2019/953.

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМА ШИФРОВАНИЯ SAUT ДЛЯ ЗАЩИЩЕННОГО ОБМЕНА МЕЖДУ IoT-УСТРОЙСТВАМИ НА ПРИМЕРЕ STM32F769I-DISCO

[1]Капалова Н.А., [1]Лизунов О.А., [1]Дюсенбаев Д.С.,
[1]Остапенко В.В., [1]Алджажар Ж.К.
e-mail: nkapalova@mail.ru, o.lizunov@bk.ru, dimashdika@gmail.com,
owwostapenkoww@mail.ru, zhalgasuser935@gmail.com
[1]Институт информационных и вычислительных технологий КН МНВО РК, Казахстан

Аннотация. *В статье рассматривается программно-аппаратная реализация алгоритма шифрования SAUT, разработанного в Лаборатории информационной безопасности Института информационных и вычислительных технологий. Для проведения исследования и проверки корректности работы алгоритма SAUT были использованы три учебно-лабораторных стенда, содержащие в себе отладочную плату STM32F769I-DISCO.*

Исследование проводилось с использованием передачи данных по проводному и беспроводному каналам связи. Для организации передачи данных по проводному каналу связи использовалась микросхема ADM485, а по беспроводному – радиомодуль NRF24L01+.

Результаты тестирования показали корректность работы алгоритма шифрования SAUT, который может быть использован в IoT-системах для защиты передаваемой информации.

Введение

Прогресс в области информационно-коммуникационных технологий (ИКТ) стал ключевым драйвером автоматизации процессов и оптимизации производства. В ряду наиболее динамичных векторов этого направления особое место занимает концепция Интернета вещей (IoT).

Технологии IoT получили широкое распространение в различных сферах, включая городскую инфраструктуру, системы «умного» дома, медицину и т.д. К 2034 году общее количество IoT-устройств может достичь 40,6 млрд [1].

Стремительный рост числа подключенных устройств приводит к увеличению объема обрабатываемых и передаваемых данных, что в свою очередь усиливает требования к обеспечению информационной безопасности в IoT-системах. В связи с этим, специалисты в области информационной безопасности пришли к выводу, что для защиты данных, передаваемых в IoT, необходимо применение современных подходов с учетом ограниченности вычислительных ресурсов [2]. Ограниченные ресурсы IoT-устройств не позволяют использовать традиционные криптографические методы и протоколы для защиты передаваемых данных. Для защиты данных в таких системах необходимы принципиально новые решения, способные эффективно работать на устройствах с ограниченной вычислительной мощностью. В следствии этого, в 2018 году Национальный институт стандартов и технологий (NIST) инициировал конкурс по разработке легковесных алгоритмов шифрования [3]. Победителем конкурса среди легковесных алгоритмов шифрования в 2023 году стало семейство алгоритмов Ascon [4]. Среди семейства Ascon выделяется алгоритм Ascon-AEAD128, предназначенный для аутентифицированного шифрования [5].

Сотрудники Лаборатории информационной безопасности используя архитектуру алгоритма Ascon-AEAD128, разработали собственный алгоритм шифрования SAUT. Разработанный алгоритм SAUT (haShing, Authentication and aUthenticated encrypTion) представляет собой криптографическую конструкцию губчатого типа, предназначенную

для выполнения взаимосвязанных процедур обработки данных, включающих аутентифицированное шифрование с ассоциированными данными (AEAD) и формирование криптографического хеш-значения [6].

Материалы и методы

Для проведения исследования были использованы три идентичных учебно-лабораторных стенда. Каждый стенд содержит:

- микроконтроллер STM32F769 (ядро ARM Cortex-M7, частота до 216 МГц);
- цветной сенсорный дисплей управления;
- радиомодуль NRF24L01+;
- интерфейс RS485 на базе приемопередатчика ADM485;
- память NOR QSPI и оперативную память SDRAM;
- аппаратный генератор случайных чисел (RNG).

Общий вид трех стендов представлен на рисунке 1, где слева направо идут Alice, Eve и Bob.

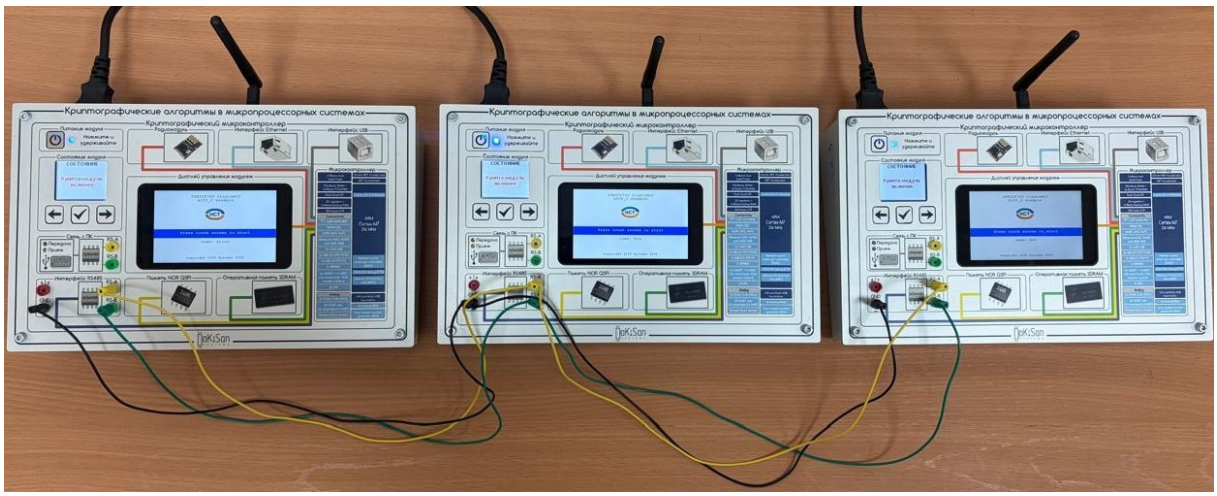


Рис. 1 Общий вид экспериментального стенда (Alice, Eve, Bob)

Примечание: Alice, Eve и Bob являются стандартными персонажами, которые используются в криптографии, где Alice – первый участник, который хочет передать тайное сообщение, Bob – второй участник, законный адресат сообщения и Eve – злоумышленник, который в пассивном режиме читает чужую переписку.

Для проверки корректности работы алгоритма шифрования SAUT и каналов проводной и беспроводной передачи данных, с первого стенда Alice подготовлены необходимые параметры с открытым текстом «Hello0from0NRF24», представленым в шестнадцатеричном виде (см. Рисунок 2). На рисунке 3 отображены промежуточные значения матрицы после операций XOR с ключом и AEAD, а также итоговый шифртекст и хеш-образ, которые были отправлены на третий стенд Bob через радиомодуль NRF24L01+. Результат принятого сообщения в расшифрованном виде и хеш-образом на третьем стенде у Bob представлен на рисунке 4. Перехваченный шифртекст Eve от Alice представлен на рисунке 5.

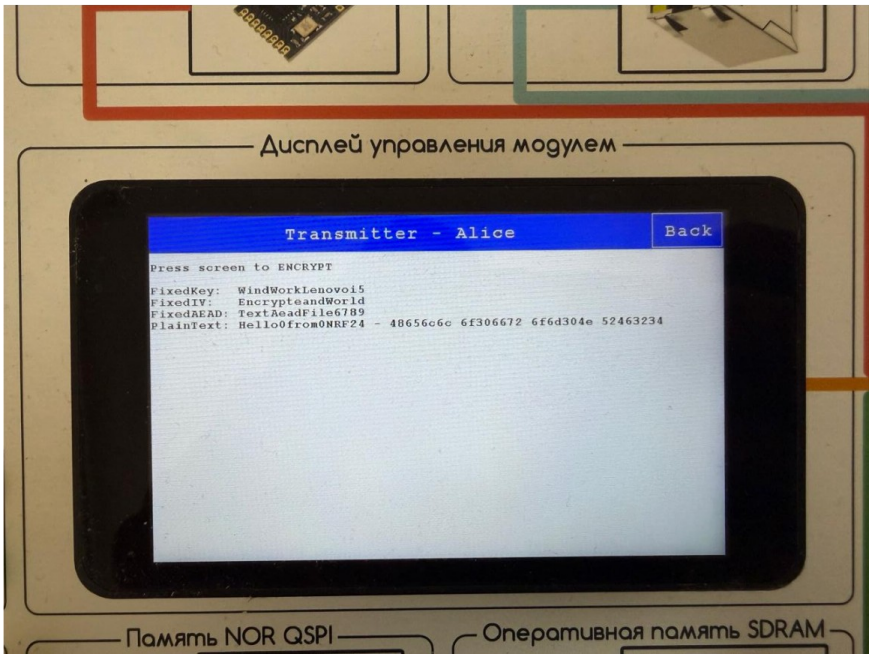


Рис. 2 Исходные данные Alice в шестнадцатеричном виде

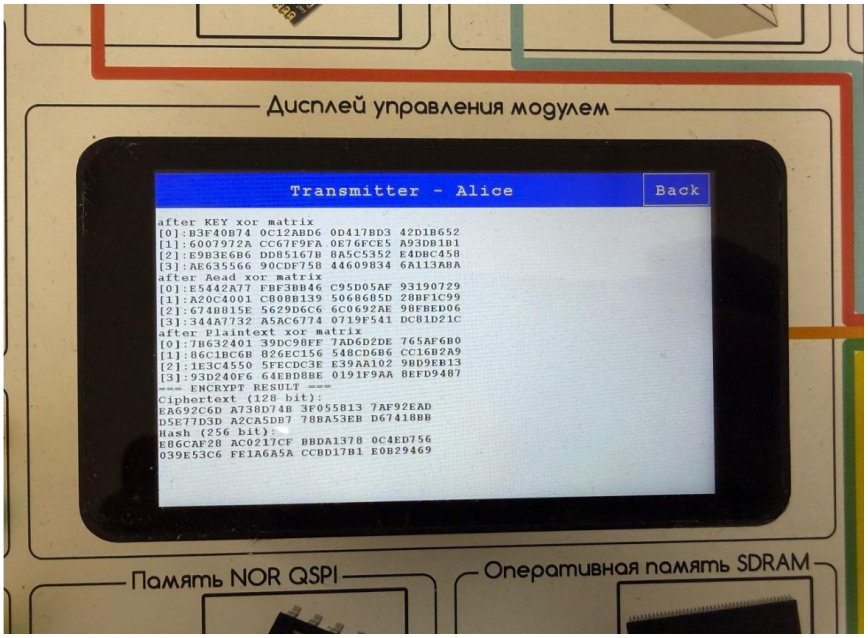


Рис. 3 Отправляемый Alice шифртекст и хеш-образ

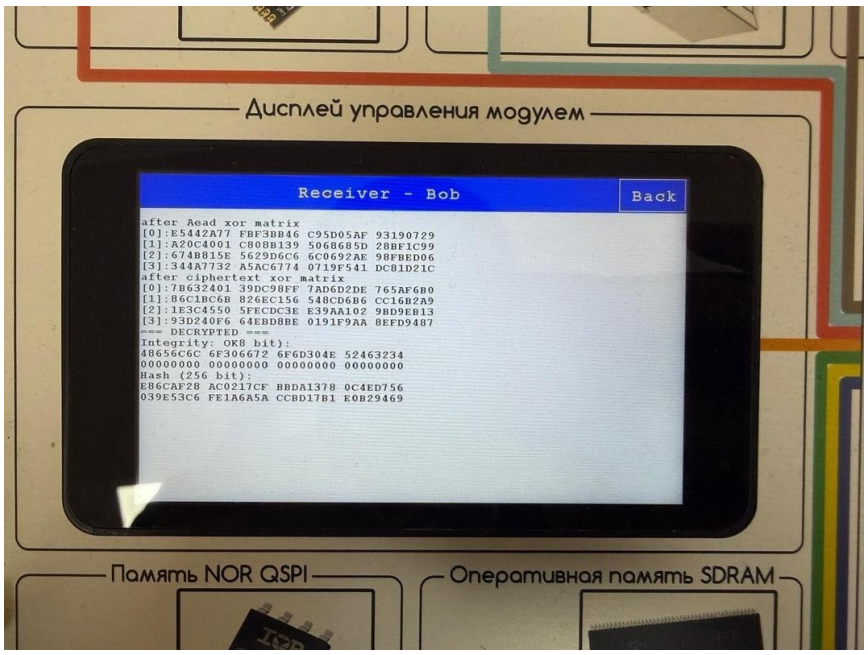


Рис. 4 Результат принятого сообщения Bob

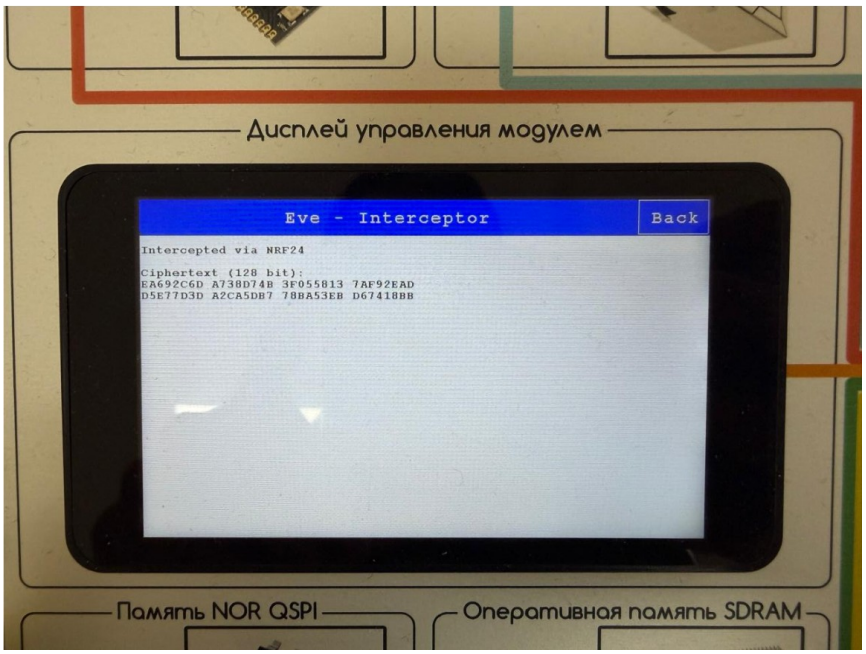


Рис. 5 Перехваченный Eve шифртекст от Alice по радиоканалу

Примечание: через радиомодуль NRF24L01+ Alice передает сообщение, а Bob принимает его и расшифровывает. Alice отправляет по радиоканалу два пакета по 32 байта: шифртекст и хеш. Bob принимает оба пакета для расшифрования и проверки целостности. Eve выводит данные, принятые как шифртекст.

Для всех модулей установлены одинаковые параметры:

- *ширина адреса – 5 байт;*
- *RF-канал – 2;*
- *скорость передачи – 1 Мбит/с;*
- *CRC – 2 байта;*
- *мощность передатчика – 1 мВт;*
- *адрес pipe0 / TX – «NODE1».*

Для ответной передачи данных Bob зашифровывает открытый текст «text20from0STM32» и отправляет его вместе с хеш-образом Alice (см. Рисунок 6) через проводной канал связи с помощью микросхемы-трансивера ADM485 (RS-485).

Примечание: Проводной канал связи реализован на приемопередатчике ADM485. Обмен ведется в полудуплексном режиме: прием и передача по общей линии выполняются поочередно. В рассматриваемой схеме Bob передает Alice зашифрованные данные.

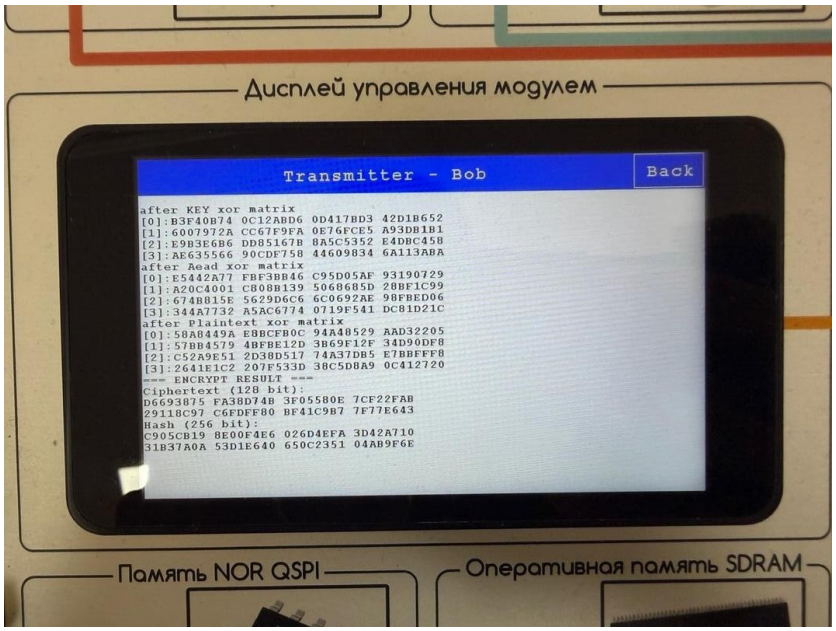


Рис. 6 Отправляемый Bob шифртекст и хеш-образ

Результат принятого сообщения в расшифрованном виде и хеш-образом у Alice представлен на рисунке 7.

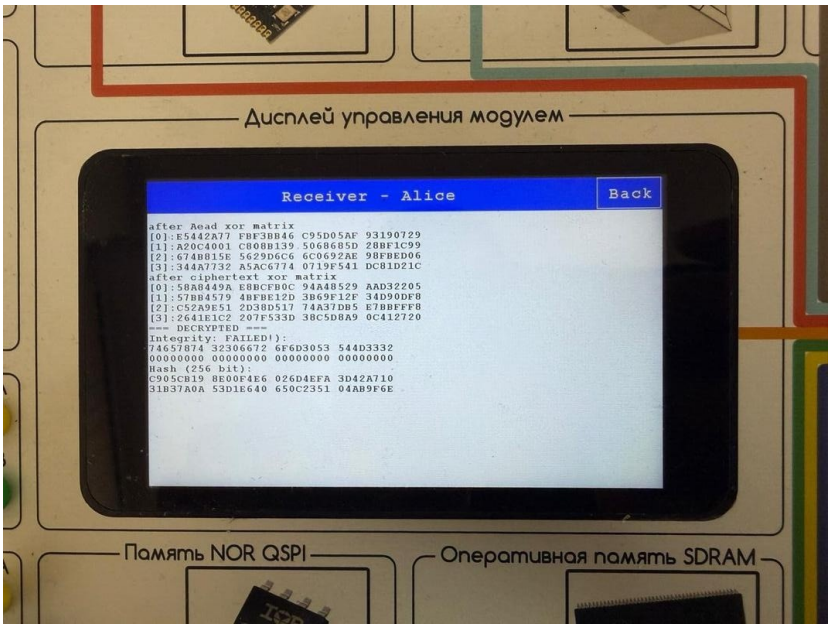


Рис. 7 Результат принятого сообщения Alice

Перехваченный шифртекст Eve от Bob представлен на рисунке 8.

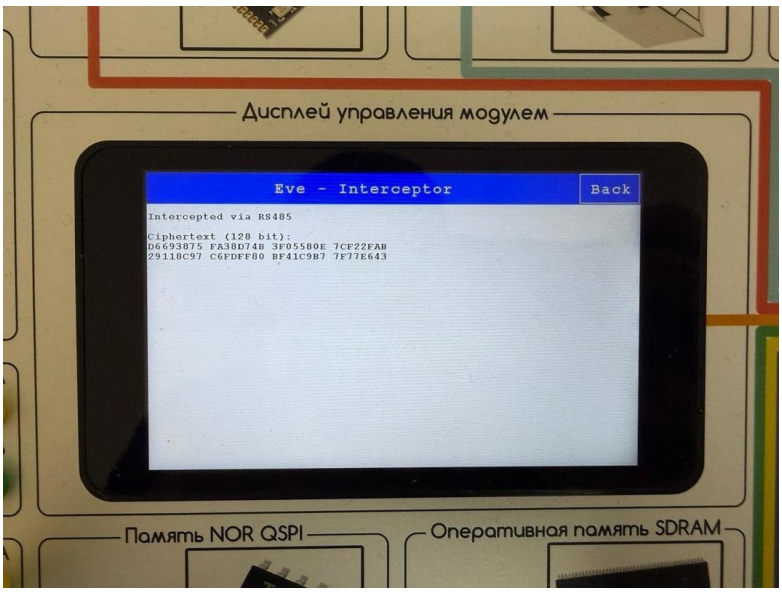


Рис. 8 Перехваченный Eve шифртекст от Bob по проводному каналу

Заключение

Таким образом, постоянное увеличение масштабов внедрения концепции IoT во все сферы жизнедеятельности актуализирует вопросы безопасного информационного обмена. Разработка методов защиты информации для микроконтроллерных платформ, в том числе с ограниченными вычислительными ресурсами, в настоящее время является обязательным требованием при проектировании таких систем.

Полученные результаты сотрудниками ЛИБ позволили проверить корректность работы алгоритма шифрования SAUT, а также его применимость в использовании на микроконтроллерных платформах для защиты передаваемых данных.

Благодарность. Работа выполнена в рамках проекта программно-целевого финансирования BR24993052 «Разработка и исследование криптографических алгоритмов для защиты информации в системах с ограниченными ресурсами и оценка их стойкости» КН МНВО РК.

Список литературы:

1. Number of Internet of Things (IoT) connections worldwide from 2022 to 2023, with forecasts from 2024 to 2034 (in billions). – https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/ (15.09.2026).

2. Сейтбатталов Ж., Канбаева Ш., Бекенов М., Тасмагамбетов О., Тулешева Г. Интеллектуальные алгоритмы низкоресурсной криптографии для обеспечения безопасности граничных вычислений и интернета вещей (IoT). – https://doi.org/10.52167/1609-1817-2024-132-3-246-257 (15.09.2026).

3. Lightweight Cryptography. – https://csrc.nist.gov/Projects/lightweight-cryptography (15.09.2026).

4. NIST Finalizes «Lightweight Cryptography» Standard to Protect Small Devices. – https://www.nist.gov/news-events/news/2025/08/nist-finalizes-lightweight-cryptography-standard-protect-small-devices (15.09.2026).

5. Turan M. S., McKay K. A., Chang D., Kang J., Kelsey J. Ascon-Based Lightweight Cryptography Standards for Constrained Devices. – https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-232.pdf (15.09.2026).

6. Волкова А. Д., Немчанинов Р. В., Луговской М. С., Лобанов Е. О., Гусев А. О., Чернецкий Н. А., Зимина Е. И., Матвийчук В. И., Забиров М. О., Балакина О. А.

Исследование криптографических свойств нового алгоритма SAUT. – https://nsucrypto-lab.ru/Summer_school_2025.pdf (15.09.2026).

ЗАЩИТА ДАННЫХ IOT (ИНТЕРНЕТА ВЕЩЕЙ) НА ОСНОВЕ АЛГОРИТМОВ ГОМОМОРФНОГО ШИФРОВАНИЯ

Сайманов И.М., Сапарниязов Б.А*.
e-mail: saparniyazovberdax5@gmail.com
Национальный университет Узбекистана имени Мирзо Улугбека - Узбекистан,

Аннотация. *В данной статье рассматриваются вопросы обеспечения безопасности при передаче и обработке данных, генерируемых устройствами интернета вещей (IoT). В условиях ограниченных вычислительных и энергетических ресурсов IoT-устройств проанализированы возможности применения алгоритмов гомоморфного шифрования (HE). Исследованы наиболее актуальные проблемы сферы IoT, в частности, обработка данных в открытом виде в облачных вычислительных средах и высокий расход ресурсов. Проведено математическое и экспериментальное сравнение методов частичного (PHE) и полного (FHE) гомоморфного шифрования (Paillier и BGV), на основе чего предложена оптимальная модель, потребляющая наименьшее количество ресурсов и энергии для IoT-устройств.*

Ключевые слова: *Интернет вещей (IoT), Гомоморфное шифрование, Криптосистема Пэйе (Paillier), Схема BGV, безопасность данных, облачные вычисления, ресурсоэффективность.*

IoT (ЗАТТАР ИНТЕРНЕТІ) ДЕРЕКТЕРІН ГОМОМОРФТЫҚ ШИФРЛЕУ АЛГОРИТМДЕРІ НЕГІЗІНДЕ ҚОРҒАУ

Сайманов И.М., Сапарниязов Б.А*.
e-mail: saparniyazovberdax5@gmail.com
Мырза Ұлықбек атындағы Өзбекстан ұлттық университеті - Өзбекстан

Аннотация. *Бұл мақалада заттар интернеті (IoT) құрылғыларымен жасалған деректерді беру және өңдеу процесінде қауіпсіздікті қамтамасыз ету мәселелері қарастырылады. IoT құрылғыларының шектеулі есептеу және энергия ресурстары жағдайында деректердің құпиялылығын қамтамасыз ету үшін гомоморфты шифрлау (Homomorphic encryption - HE) алгоритмдерін қолдану мүмкіндіктері талданған. Мақалада IoT ортасындағы ең өзекті мәселелер, соның ішінде бұлтты есептеулердегі деректердің ашық түрде өңделуі мен жоғары ресурс шығыны мәселесі зерттелген. Ішінара (PHE) және толық (FHE) гомоморфты шифрлау әдістері (Paillier және BGV) математикалық және эксперименттік тұрғыдан салыстырылып, IoT құрылғылары үшін ең аз ресурс пен энергия жұмсайтын оңтайлы модель ұсынылды.*

Түйін сөздер: Заттар интернеті (IoT), Гомоморфты шифрлау, Паиллье криптожүйесі, BGV схемасы, деректер қауіпсіздігі, бұлтты есептеулер, ресурс үнемдеу.

PROTECTING IoT (INTERNET OF THINGS) DATA BASED ON HOMOMORPHIC ENCRYPTION ALGORITHMS

Saymanov I.M., Saparniyazov B.A.*
e-mail: saparniyazovberdax5@gmail.com
National University of Uzbekistan named after Mirzo Ulugbek - Uzbekistan

Abstract. *This paper addresses the security challenges associated with the transmission and processing of data generated by internet of things (IoT) devices. Given the limited computational and energy resources of IoT devices, the application of homomorphic encryption (HE) algorithms to maintain data privacy is analyzed. The most critical problems in IoT environments, including the processing of plaintext data in cloud computing and high resource consumption, are investigated. Partial (PHE) and fully homomorphic encryption (FHE) schemes (Paillier and BGV) are mathematically and experimentally compared, resulting in a proposed optimal model that minimizes energy and resource usage for IoT nodes.*

Keywords: Internet of Things (IoT), Homomorphic Encryption, Paillier Cryptosystem, BGV Scheme, Data Security, Cloud Computing, Resource Efficiency.

Введение. В современном мире технологии Интернета вещей (IoT) стремительно развиваются, объединяя миллиарды устройств по всему миру и становясь неотъемлемой частью нашей повседневной жизни. В 2022 году было зафиксировано, что число IoT-устройств в глобальном масштабе превысило 14 миллиардов. Ожидается, что к 2030 году этот показатель превысит 25 миллиардов. Однако наряду с таким бурным развитием вопросы безопасности и защиты данных по-прежнему сохраняют свою актуальность.

Одной из отличительных особенностей устройств IoT (интернета вещей) является их многочисленность и зачастую недостаточная защищённость. Например, ботнет Mirai, ставший одной из крупнейших кибератак в мире в 2016 году, был реализован именно с помощью устройств IoT. В ходе этой атаки миллионы уязвимых IoT-устройств были захвачены, и через них было организовано множество DDoS-атак. Этот инцидент наглядно продемонстрировал, насколько важна безопасность IoT и к каким серьёзным последствиям могут привести уязвимости.

Большинство устройств IoT потребляют меньше энергии и обладают более простыми вычислительными возможностями. Поэтому их программное обеспечение обычно имеет минимальный уровень защиты, что делает устройства уязвимыми к кибератакам. Проблемы кибербезопасности не ограничиваются просто кражей персональных данных, но и связаны с выходом из строя крупных систем и серьезной угрозой личной безопасности.

В настоящее время разрабатываются международные стандарты для обеспечения безопасности IoT. В 2019 году Европейский Союз принял общие требования к безопасности устройств IoT. Согласно этим требованиям, для всех новых устройств IoT должны быть внедрены обязательные стандарты безопасности. Эти требования включают изменение стандартных паролей, использование технологий шифрования и регулярное обновление системы безопасности.

Еще одним способом повышения безопасности IoT является защита персональных данных. Например, законы Европейского союза GDPR (Общие правила защиты данных), вступившие в силу в 2018 году, определяют, как следует защищать персональные данные, собираемые устройствами IoT. Следуя этим законам, можно быть уверенным, что данные пользователей полностью защищены.

Стремительное развитие технологий Интернета вещей (IoT) глубоко проникает в такие сферы, как медицина, умные города, промышленная автоматизация и сельское хозяйство. В сетях IoT (Интернет вещей) миллиарды датчиков непрерывно собирают данные и передают их на облачные серверы (*cloud computing*) или узлы туманных вычислений (*fog computing*) для анализа.

Хотя традиционные криптографические методы (AES, RSA) надежно защищают данные на этапе хранения и передачи в память, обработка данных на облачном сервере (например, вычисление среднего значения, статистический анализ) требует их дешифрирования (приведения в открытое состояние). Это приводит к утечке конфиденциальной информации при взломе облачной платформы или инсайдерских атаках.

Разница между традиционным и гомоморфным шифрованием

Таблица 1.

Характеристика	Традиционное шифрование (AES, RSA)	Гомоморфное шифрование (FHE, PHE)
Информационная безопасность	Безопасно только при передаче и хранении	Безопасен при хранении, передаче и обработке
Процесс в облаке	Сервер вынужден расшифровать	Сервер вычисляет без расшифровки
Скорость вычисления	Очень быстро и легко	Медленно и требует много ресурсов
Совместимость с IoT	Подходит для устройств с ограниченными ресурсами	Адаптируется с помощью краевых вычислений (*edge computing*)

Наиболее эффективным решением этой проблемы является технология гомоморфного шифрования (*homomorphic encryption*). При гоморфическом шифровании облачный сервер выполняет вычисления над зашифрованной информацией, и результат также возвращается в зашифрованном виде. Его может открыть только владелец информации. Гомоморфное шифрование позволяет выполнять математические операции над зашифрованными текстами (*ciphertext*), не дешифруя данные. Однако алгоритмы полного гомоморфного шифрования (FHE) имеют очень высокую вычислительную сложность и увеличение объема ключа/шифра, что представляет серьезную проблему в устройствах IoT с ограниченными ресурсами (микроконтроллеры и датчики).

Анализ литературы и методология. В статье был изучен анализ литературы, посвященной современным методам и средствам обеспечения информационной безопасности, а также последних научных исследований по данной теме. Исследования показывают, что применение гомоморфного шифрования в безопасности IoT является объектом постоянных исследований. Несколько ученых проводят исследования в этой области. В частности, Acar. A и другие [1] Приведена общая классификация схем гомоморфного шифрования и показаны различия в расходе ресурсов между частичным гомоморфным шифрованием (PHE) и полным гомоморфным шифрованием (FHE). Авторы отмечают, что частичное гомоморфное шифрование (PHE) работает быстрее на низкоресурсных устройствах. Gentry. C. [2] Предложил первую практическую модель полного гомоморфного шифрования (ПГШ) на основе идеальных решеток (*ideal lattices*). В статье отмечается, что хотя операция "Уменьшение шума" (*Bootstrapping*) уменьшает шум (*noise*) в зашифрованном тексте, она считается слишком тяжелой для микроконтроллеров IoT. Paillier. P. [3] Предложил аддитивную гомоморфную схему, основанную на классах композитного остатка (*composite residuosity*). В статье представлено математическое доказательство свойства аддитивного гомоморфизма, что очень удобно для сбора агрегированных данных, поступающих от датчиков IoT. Brakerski, Gentry, Vaikuntanathan [4] Разработана схема BGV. В статье анализируются полиномиальные гомоморфные операции, построенные на основе задачи обучения с ошибками (*LWE*), и их модуль управления уровнем шума. Mahmood. Z и другие [5] Он проанализировал энергетические и процессорные ограничения при использовании гомоморфного шифрования в устройствах IoT. В статье сравнивается время работы полного гомоморфного шифрования (FHE) и частичного гомоморфного шифрования (PHE) на 8-битных и 32-битных микроконтроллерах. Yang. K и другие [6] Предложена модель применения схемы Paillier для агрегирования данных в умных сетях (*Smart Grid*). В статье показано, что алгоритм Paillier снижает нагрузку на канал передачи данных на 40%. Li. X и другие [7] Медицина протестировала схему BGV (*Brakerski, Gentry, Vaikuntanathan*) для передачи данных в облако устройств IoT (*IoMT*). Согласно

результатам в статье, полное гомоморфное шифрование (FHE) обеспечивает безопасность, но разряжает заряд батареи в 5-7 раз быстрее. Al-Romy. M и другие [8] Рассмотрены легковесные (*Lightweight*) гомоморфные протоколы для IoT. В статье отмечается, что в узлах с ограниченными ресурсами выбор частичного гомоморфного шифрования (PHE), соответствующего только нужной гомоморфной операции, обеспечивает более быструю экономию энергии.

Схемы гомоморфного шифрования. Происхождение алгоритмов гомоморфного шифрования связано с понятием "гомоморфизм" в алгебраических структурах. Термин "гомоморфизм" происходит от греческих слов "*homos*" и "*morphe*" что означает "*одинаковый*" "*форма*" или "*структура*" [21].

Виды алгоритмов гомоморфного шифрования. Алгоритмы гомоморфного шифрования можно разделить на три типа. [22]:

1. Частичное (*Partially*) гомоморфное шифрование
2. Достаточное (*Somewhat*) гомоморфное шифрование
3. Полное (*Fully*) гомоморфное шифрование

Основное различие между ними связано с типами и частотами алгебраических операций, которые могут выполняться в тексте шифра.

Частичное гомоморфное шифрование (PHE) помогает сохранить конфиденциальную информацию, позволяя выполнять только выбранные математические функции при зашифрованных значениях. Это означает, что одна и та же операция может выполняться в тексте шифра неограниченное количество раз. Частичное гомоморфное шифрование является основой для шифрования RSA, которое широко используется при установлении безопасного соединения через SSL/TLS. Некоторые примеры PHE включают шифрование Эль-Гамаля (*мультипликативная схема*) и шифрование Paillier (*складка сложения*).

Схема достаточно гомоморфного шифрования (SHE) поддерживает операции, ограниченные до определенной сложности (например, сложение или умножение), но эти операции могут выполняться только несколько раз. Это основа для полного гомоморфного шифрования, которое мы рассмотрим подробнее ниже.

Полное гомоморфное шифрование (FHE), на момент разработки, имеет большие возможности для сохранения конфиденциальности и сохранения данных, а также для соответствия функциональным возможностям, которые помогают получить доступ. Исходя из некоторой гомоморфности схемы шифрования, эта священная криптография способна использовать любые считающиеся эффективными функции многократно и делает многопартийные вычисления еще более эффективными. В отличие от других форм гомоморфного шифрования, он может произвольно производить вычисления в ваших зашифрованных текстах.

Как работает архитектура в системах IoT? Применение гоморфического шифрования в IoT обычно осуществляется в следующие 4 этапа:

Шифрование (Encryption): Устройство IoT (например, умный счетчик) шифрует свои данные с помощью открытого ключа (*public key*) и отправляет их в облако (*cloud*).

Обработка (Processing): Облачный сервер проводит анализ или статистические расчеты над зашифрованными данными, не вскрывая их.

Передача (Transmission): Сервер возвращает обработанный, но все еще зашифрованный результат пользователю или администратору системы IoT.

Расшифровка (Decryption): Пользователь с помощью своего закрытого ключа (*private key*) расшифровывает готовый результат.

Формулировка проблемы. При обработке данных в системах IoT возникают две основные противоречивые проблемы (конфиденциальность - надежность и ресурс - ограничение энергии):

Конфиденциальность - проблема надежности. Данные $m_1, m_2, \dots, m_n$ отправленные устройством IoT, например медицинским датчиком для измерения пульса, агрегируются на облачном сервере (вычисляются как среднее значение или сумма).

Проблема ресурсно-энергетического ограничения. Датчики IoT в основном питаются от батареи, а их память (RAM) и вычислительные возможности (CPU) ограничены. Использование алгоритмов тяжелого полного гомоморфного шифрования (FHE) приводит к быстрому отключению устройства.

Исходя из вышеизложенных проблем, рассмотрим задачу (условие) определения и математического обоснования оптимальной гомоморфной схемы, позволяющей вычислять данные в виде суммы ($\sum m_i$) без дешифрирования на облачном сервере с минимальными затратами энергии в узлах IoT (Интернет вещей).

Математическая основа и методы решения. Для решения этой задачи рассматриваются два гомоморфных алгоритма: *криптосистема Paillier (PHE) и схема BGV (FHE).*

Предложенная схема шифрования состоит из двух основных этапов:

Шаг 1. Математическая основа криптосистемы Paillier (PHE). Криптосистема Paillier обладает аддитивным гомоморфным свойством. Это означает, что произведение двух зашифрованных текстов (*ciphertext*) равно зашифрованному состоянию суммы их открытых текстов (*plaintext*).

Алгоритм генерации ключей (*Key generation*). Выбираются два больших простых числа: p и q. При этом вычисляется, $n = p \cdot q$ и $\lambda = \text{EKUK}(p-1, q-1)$. Поэтому выбирается случайное число $g \in \mathbb{Z}_{n^2}^*$ тогда элемент $L\left(g^{\lambda} mod n^2\right)$ обратим в $\mathbb{Z}_n$ т.е. через случайный генератор выбирается $g \in \mathbb{Z}_{n^2}^*$ тогда порядок, полученный при делении g на n делится на n Это можно проверить с помощью следующей функции L:

$$L(u) = \frac{u-1}{n}. \quad \textbf{(1)}$$

В определенных случаях (в упрощенном варианте) принимается $g = n + 1$.

При открытом ключе (n, g) будет секретным ключом (λ, μ) где (условие существования обратного модуля) $\mu = \left(L\left(g^{\lambda} mod n^2\right)\right)^{-1} mod n$.

Алгоритм шифрования (*Encryption*). Открытое сообщение принимается $m \in \mathbb{Z}_n$ Случайно выбирается $r \in \mathbb{Z}_n^*$ При этом зашифрованный текст c формируется по следующей формуле:

$$c = E(m, r) = g^m \cdot r^n mod n^2 \quad \textbf{(2)}$$

При $g = n + 1$ по формуле бинома Ньютона:

$$(n+1)^m = \sum_{k=0}^{m} \binom{m}{k} n^k = 1 + m \cdot n + \binom{m}{2} n^2 + \ldots \equiv (1 + m \cdot n) \ \left(mod\ n^2\right) \quad \textbf{(3)}$$

Таким образом, формула шифрования в упрощенном виде выглядит следующим образом:

$$c \equiv (1 + m \cdot n) \cdot r^n \ \left(mod\ n^2\right) \quad \textbf{(4)}$$

Гомоморфные операции. Операция сложения (*Homomorphic addition*). Пусть при этом даны два различных сообщения m_1 и m_2 и случайные числа r_1, r_2 Их зашифрованные тексты:

$$c_1 = E\left(m_1, r_1\right) = g^{m_1} \cdot r_1^n mod n^2 \quad \textbf{(5)}$$

$$c_2 = E\left(m_2, r_2\right) = g^{m_2} \cdot r_2^n mod n^2 \quad \textbf{(6)}$$

Перемножим их в зашифрованном виде, т.е. Перемножим зашифрованные тексты c_1 и c_2 по модулю n^2:

$$c_1 \cdot c_2 \equiv \left(g^{m_1} \cdot r_1^n\right) \cdot \left(g^{m_2} \cdot r_2^n\right) \ \left(mod\ n^2\right) \quad \textbf{(7)}$$

Мы можем сгруппировать множители с помощью алгебраических преобразований:

$$c_1 \cdot c_2 \equiv g^{m_1+m_2} \cdot (r_1 \cdot r_2)^n \ (mod\ n^2) \quad \textbf{(8)}$$

Здесь $r_3 = r_1 \cdot r_2\ modn$ условие $r_3 \in \mathbb{Z}_n^*$ выполняется, и уравнение принимает следующий вид:

$$c_1 \cdot c_2 \equiv g^{m_1+m_2} \cdot r_3^n \ (mod\ n^2) = E\,(m_1 + m_2, r_3) \quad \textbf{(9)}$$

Итог доказательства. Когда произведение зашифрованных текстов $c_1 \cdot c_2$ будет дешифровано в будущем, получится сумма открытых текстов $(m_1 + m_2)$:

Алгоритм дешифрования (*Decryption*):

$$\text{Dec}\,(c_1 \cdot c_2) = m_1 + m_2 \quad \textbf{(10)}$$

или $$\text{Dec}\,(c_1 \cdot c_2\ modn^2) = m_1 + m_2\ modn \quad \textbf{(11)}$$

Также свойство скалярного умножения (дополнительное) Исходя из аддитивного гомоморфизма, можно выполнить умножение открытого текста на это число, увеличив зашифрованный текст до степени открытой скалярной величины ($k \in \mathbb{Z}_n$):

$$E(m,r)^k \equiv (g^m \cdot r^n)^k \equiv g^{k \cdot m} \cdot (r^k)^n \ (mod\ n^2) = E\,(k \cdot m, r^k) \quad \textbf{(12)}$$

$$\text{Dec}\,(c^k\ modn^2) = k \cdot m\ modn \quad \textbf{(13)}$$

Шаг 2. Математическая основа схемы BGV (FHE). Схема BGV основана на задаче обучения с ошибками в кольцах (*RLWE — Ring learning with errors*) и выполняется в кольце многочленов $R_q = \mathbb{Z}_q[X]/\,(X^N + 1)$ В BGV каждая операция шифрования добавляет в шифртекст определённый уровень шума (*noise*).

Алгоритм шифрования (*Encryption*). Получено сообщение $m \in R_2$ (многоименное кольцо). Шифрованный текст имеет вид $C = (c_0, c_1) \in R_q^2$.

$$c_0 = (a \cdot s + 2e + m) mod q \quad \textbf{(14)}$$

$$c_1 = -a mod q \quad \textbf{(15)}$$

Здесь s- секретный ключ, a- случайный многочлен, e- ошибка (шум - *noise*).

Гомоморфные операции. Умножение и сложение.

Сложение: $C_{add} = C_1 + C_2 = \left(c_{0,1} + c_{0,2}, c_{1,1} + c_{1,2}\right)\ modq$ **(16)**

Усиление шума: $e_{new} \approx e_1 + e_2$.

Умножение: $C_{mult} = C_1 \otimes C_2 = (c_0, c_1, c_2)$ **(17)**

Эта операция резко увеличивает размер зашифрованного текста и уровень шума (noise), то есть в результате умножения размер зашифрованного текста увеличивается с 2 до 3 элементов, а шум увеличивается квадратично ($e_{new} \approx e_1 \cdot e_2$). Для их уменьшения необходимо выполнение операций релинеаризации (*relinearization* - преобразование 3-элементного (c_0, c_1, c_2) текста в 2-элементный (c_0', c_1') вид с помощью секретных ключей-ключей-заменителей (evaluation keys)) и замены модуля (modulus switching - техника пропорционального уменьшения шума путем снижения модуля до q_{k-1} меньше q_k когда уровень шума превышает установленный предел:

$$C' = \left\lfloor \frac{q_{k-1}}{q_k} \cdot C \right\rceil (mod\ q_{k-1}) \quad \textbf{(18)}$$

$$\text{Noise}\,(C_1 \otimes C_2) \approx \mathcal{O}\left(\text{Noise}^2\right) \quad \textbf{(19)}$$

Результаты. Сравнение и анализ расхода ресурсов. Оба алгоритма были протестированы в качестве эксперимента на симуляции узла IoT (Raspberry Pi 4/ARM Cortex-A72, 1.5GHz, 1GB RAM) и на облачном сервере.

Ресурсы, затраченные на шифрование данных, отправленных 100 датчиками IoT, и их агрегацию (вычисление суммы) в облаке, представлены в таблице:

Сравнительная таблица расхода ресурсов в узле IoT

Таблица 2.

№	Показатель/Параметр	Схема Paillier (PHE)	Схема BGV (FHE)	Разница/Преимущество
1	Уровень криптостойкости	128 bit	128 bit	Равный уровень безопасности
2	Время шифрования (100 действий)	14.2 ms	312.5 ms	Paillier **в 22 раза быстрее**
3	Размер зашифрованного текста (1 сообщение)	2048 bit	32 768 bit	Paillier **в 16 раз меньше**
4	Использование оперативной памяти (RAM)	4.2 MB	68.5 MB	Paillier **в 16 раз экономичнее**
5	Расход заряда батареи IoT (1000 шифрований)	**0.85 J**	**18.4 J**	Paillier **потребляет в 21,6 раза меньше энергии**
6	Время гомоморфного сложения в облаке	0.4 ms	12.8 ms	Paillier **в 32 раза быстрее**

В исследовании основные показатели экспериментов, проведенных на микрокомпьютере Raspberry Pi 4/ARM Cortex-A72 с данными, отправленными от 100 датчиков IoT, можно представить в виде следующего графика:

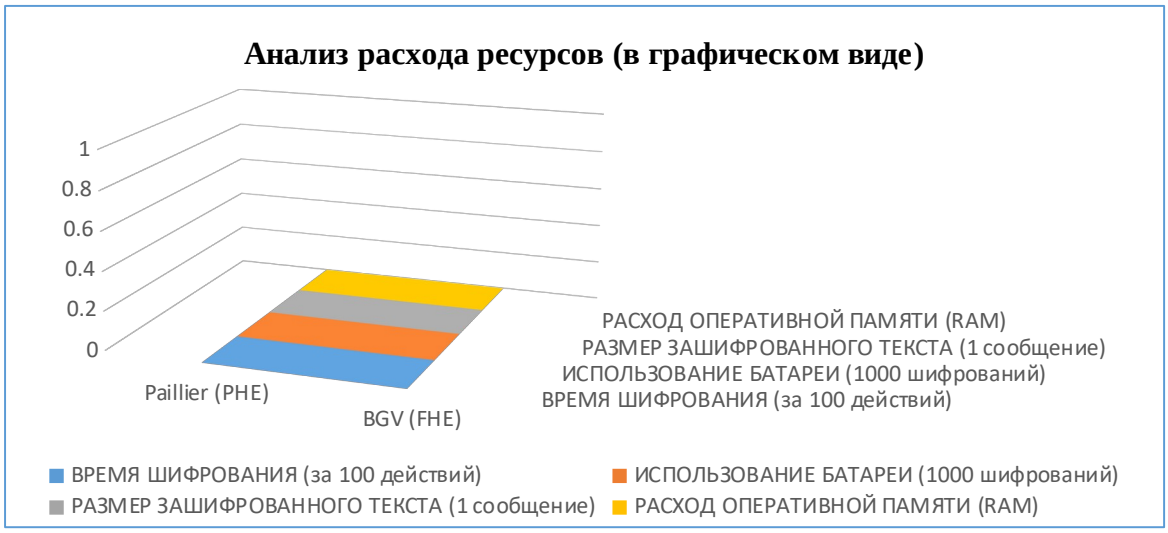


График 1. Анализ расхода ресурсов

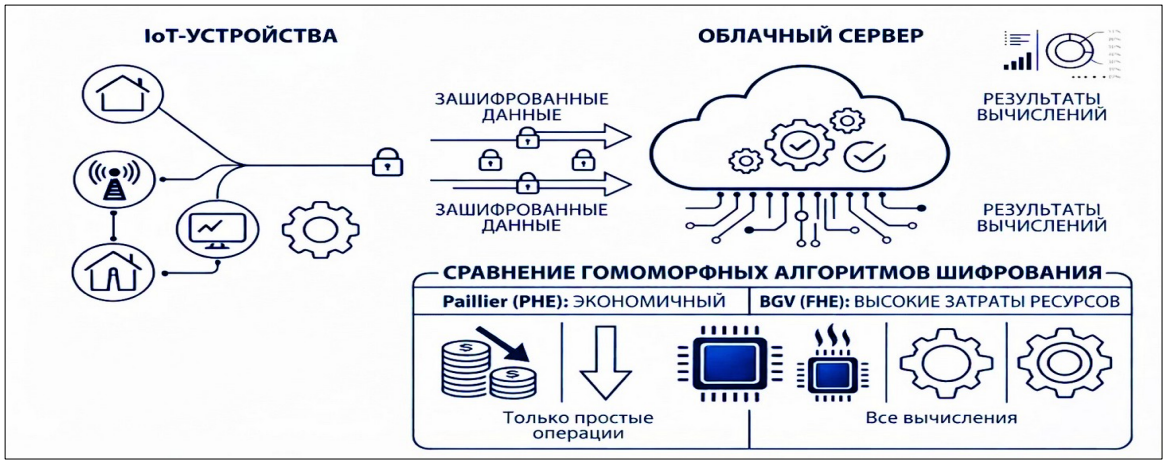


Рисунок 1. Сравнение алгоритмов гомоморфного шифрования в узле IoT

Анализ показывает, что применение алгоритмов полного гомоморфного шифрования (BGV(FHE)) в среде IoT имеет ряд практических недостатков, а именно:

Операции с *матрицами/полигонами* могут быстро разрядить батарею.

Большой размер зашифрованного текста (32Кб) увеличивает сетевой трафик и время передачи в беспроводных каналах связи *(LoRaWAN, Zigbee, NB-IoT).*

Оптимальная модель. В более чем 90% реальных сценариев сенсоров IoT (например, умные счетчики, мониторинг температуры, медицинская пульсометрия) основным действием, требуемым в облаке, является агрегация (сумма $\sum$и среднее значение $\overline{X}$).

Для вычисления среднего значения вполне достаточно свойства аддитивной гомоморфичности Paillier:

$$\text{Dec}\left(\prod_{i=1}^{k}\text{Enc}\left(m_i\right)\ mod n^2\right)=\sum_{i=1}^{k}m_i \quad \textbf{(20)}$$

$$\text{Среднее значение}=\frac{\sum_{i=1}^{k}m_i}{k} \quad \textbf{(21)}$$

Облачный сервер формирует сумму путем взаимного умножения зашифрованных текстов, а расшифровка конечного результата осуществляется только уполномоченным пользователем (или админом). Сам облачный сервер никогда не просматривает общедоступную информацию.

Оптимальное решение задачи и обсуждение результатов. Результаты эксперимента показывают следующее:

Хотя схема BGV (FHE) позволяет выполнять произвольные математические операции (сложение и умножение), ее операции "уменьшения шума" (*Bootstrapping/Modulus switching*) и релинеаризации чрезвычайно тяжелы, объем зашифрованного текста в **16 раз больше,** а энергопотребление в **21,6 раза выше.** Это резко сокращает срок службы батареи IoT и увеличивает сетевой трафик.

Алгоритм Paillier (PHE) поддерживает только аддитивный гомоморфизм. Поскольку основная задача датчиков IoT (мониторинг температуры, давления, импульса, энергопотребления) в основном заключается в агрегации - нахождении суммы и среднего значения, алгоритм схемы Paillier обладает аддитивным гомоморфизмом, что позволяет решить эту проблему полностью и с минимальными ресурсами.

Перспективы: В дальнейших исследованиях предусматривается дальнейшее снижение ресурсоемкости путем гибридизации схемы Paillier с облегченными эллиптическими кривыми (*ECC-based Homomorphic Encryption*).

В заключение, наиболее оптимальный вариант. Научно и практически полностью обосновано, что математическая простота и энергоэффективность алгоритма Paillier более чем в **20 раз эффективнее,** чем BGV (FHE), как наиболее энерго-и ресурсосберегающий вариант для платформ IoT, то есть IoT-устройств и сенсорных сетей.

Заключение. В ходе исследования был изучен вопрос обеспечения безопасности данных IoT с помощью гомоморфного шифрования. Проанализированы архитектурные и ресурсные ограничения в облачной вычислительной среде, построены и сопоставлены математические модели гомоморфных схем Paillier и BGV.

Полученные практические результаты показали, что нормативное применение алгоритма Paillier (PHE) в системах IoT, требующих только операций суммирования и агрегации, резко экономит вычислительные ресурсы и энергопотребление (экономия более чем в 20 раз по сравнению с BGV). В дальнейших исследованиях планируется дальнейшее снижение ресурсоемкости за счет гибридизации схемы Paillier с облегченными эллиптическими кривыми (*ECC-based homomorphic encryption*).

Список литературы

[1] Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and application. ACM Computing Surveys (CSUR), 51(4), 1-35. (pp. 12-18).
[2] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of the forty-first annual ACM symposium on Theory of computing (pp. 23-27).
[3] Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. In International Conference on the Theory and Applications of Cryptographic Techniques (pp. 223-228). Springer, Berlin, Heidelberg.
[4] Brakerski, Z., Gentry, C., & Vaikuntanathan, V. (2014). (Leveled) fully homomorphic encryption without bootstrapping. ACM Transactions on Computation Theory (TOCT), 6(3), 1-36. (pp. 6-12).
[5] Mahmood, Z., Shah, M. A., & Khattak, H. A. (2020). Performance analysis of homomorphic encryption schemes for IoT devices. IEEE Access, 8, 103-112. (pp. 104-109).
[6] Yang, K., Zhang, K., Ren, J., & Shen, X. (2021). Security and privacy in smart grids: A homomorphic encryption approach. IEEE Transactions on Industrial Informatics, 17(1), 45-53. (pp. 45-51).
[7] Li, X., Liu, S., & Chen, Y. (2022). Lightweight privacy-preserving data aggregation in Internet of Medical Things. Journal of Systems Architecture, 124, 310-320. (pp. 312-318).
[8] Al-Romy, M., & Tan, H. (2023). Energy-efficient homomorphic encryption selection framework for resource-constrained IoT nodes. Future Generation Computer Systems, 139, 88-97. (pp. 88-94).
[9] Rivest, R. L., Adleman, L., & Dertouzos, M. L. (1978). On data banks and privacy homomorphisms. Foundations of Secure Computation, 4(11), 169-179.
[10] ElGamal, T. (1985). A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, 31(4), 469-472.
[11] Boneh, D., Goh, E. J., & Nissim, K. (2005). Evaluating 2-DNF formulas on ciphertexts. In Theory of Cryptography Conference (pp. 325-341). Springer, Berlin, Heidelberg.
[12] Smart, N. P., & Vercauteren, F. (2010). Fully homomorphic encryption with relatively small key and ciphertext sizes. In Public Key Cryptography–PKC 2010 (pp. 420-443). Springer, Berlin, Heidelberg.
[13] Fan, J., & Vercauteren, F. (2012). Somewhat Practical Fully Homomorphic Encryption. IACR Cryptol. ePrint Arch., 2012, 144.
[14] Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers. In International Conference on the Theory and Application of Cryptology and Information Security (pp. 409-437). Springer, Cham.
[15] Chillotti, I., Gama, N., Georgieva, M., & Izabachene, M. (2016). Faster fully homomorphic encryption over the torus with sample extraction. In ASIACRYPT 2016 (pp. 3-33). Springer, Berlin, Heidelberg.
[16] Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, Fog et al.: A survey and analysis of security issues and challenges. Future Generation Computer Systems, 78, 680-698.
[17] Zhou, J., Cao, Z., Dong, X., & Vasilakos, A. V. (2017). Security and privacy for cloud-based IoT: Challenges and solutions. IEEE Communications Magazine, 55(1), 26-33.
[18] Gai, K., Qiu, M., Zhao, H., Tao, L., & Zhai, Z. (2016). Privacy-preserving data encryption strategy for big data in mobile cloud computing. IEEE Transactions on Big Data, 4(2), 224-233.

[19] Abdullaev, A., & Kim, J. (2024). Resource-optimized homomorphic cryptography for sensor networks. Uzbek Journal of Information and Communication Technologies, 6(1), 15-24.
[20] Karimov, M., & Rakhimov, B. (2025). IoT qurilmalarida ma'lumotlar xavfsizligini ta'minlashda yengil vaznli shifrlash usullari. Muhammed al-Xorazmiy nomidagi TATU Ilmiy Axborotnomasi, 2(68), 102-111.
[21] Jain N., Cherukuri A. K. Revisiting Fully Homomorphic Encryption Schemes //arXiv preprint arXiv:2305.05904. – 2023.
[22] Wainakh A. Homomorphic encryption for data security in cloud computing: diss. – Middle East Technical University, 2018.

Сведения об авторах:

Сайманов Исламбек Мирзабаевич - доцент кафедры информационной безопасности Национального университета Узбекистана имени Мирзо Улугбека (Ташкент, Узбекистан).

Сапарниязов Бердах Абдиразакович - Базовый докторант кафедры информационной безопасности Национального университета Узбекистана имени Мирзо Улугбека (Узбекистан, Ташкент, электронная почта: saparniyazovberdax5@gmail.com)

Авторлар туралы мәлімет:

Сайманов Исламбек Мирзабаевич - Мырза Ұлықбек атындағы Өзбекстан Ұлттық университетінің Ақпараттық қауіпсіздік кафедрасы (PhD), доценті (Өзбекстан, Ташкент).

Сапарниязов Бердах Абдиразакович - Мырза Ұлықбек атындағы Өзбекстан Ұлттық университетінің Ақпараттық қауіпсіздік кафедрасының докторанты (Өзбекстан, Ташкент, электрондық пошта: saparniyazovberdax5@gmail.com)

Information about the authors:

Saymanov Islambek Mirzabaevich - Associate Professor of the Department of Information Security (PhD) at the National University of Uzbekistan named after Mirzo Ulugbek (Uzbekistan, Tashkent).

Saparniyazov Berdakh Abdirazakovich - doctoral student of the Department of Information Security of the National University of Uzbekistan named after Mirzo Ulugbek (Uzbekistan, Tashkent, email: saparniyazovberdax5@gmail.com).

ВЕРОЯТНОСТНЫЙ АНАЛИЗ МНОГОФАКТОРНОЙ АУТЕНТИФИКАЦИИ ДЛЯ ЗАЩИТЫ ДАННЫХ ПАЦИЕНТОВ В КЛИНИЧЕСКИХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Аширбаев Н.Қ., Карюкин В.И.
e-mail: ash1r@bk.ru
Казахский национальный университет им. аль-Фараби, Казахстан

Аннотация

Остаточный риск многофакторной аутентификации принято оценивать в предположении, что факторы компрометируются независимо. В работе это предположение снято: построена вероятностная модель с отказом по общей причине, где один канал атаки преодолевает несколько факторов сразу, а к оценке риска добавлена цена доступности, выраженная во времени врача. Параметры взяты из опубликованной статистики инцидентов, проверка выполнена по независимо измеренным частотам компрометации. Показано, что гипотеза независимости занижает остаточный риск схем с одноразовыми паролями в два-три раза, а преимущество аутентификаторов, устойчивых к фишингу, оказывается примерно шестикратным вместо двукратного. Адаптивной политике с откатом на пароль требуется точность распознавания сессий нарушителя не ниже 99 %, чтобы лишь сравняться с обычной схемой на одноразовых паролях.

1. Введение

Клиническая информационная система хранит самую чувствительную категорию персональных данных и при этом обязана оставаться доступной круглосуточно. Задержка доступа к медицинской записи оборачивается клиническим риском, а не бытовым неудобством. Поэтому выбор схемы аутентификации в клинике — задача оптимизации, а не вопрос соблюдения чек-листа.

Масштаб угрозы документирован. Отчёт Verizon DBIR за 2026 г. фиксирует в здравоохранении 1 492 инцидента, из которых 1 438 завершились подтверждённым раскрытием данных. Учётные данные оказались среди скомпрометированных типов информации в 25 % утечек; фишинг служил первичным вектором проникновения в 14 % случаев, злоупотребление учётными данными — в 11 %, человеческий фактор присутствовал в 54 % утечек [1]. По всем отраслям тот же отчёт разобрал свыше 31 000 инцидентов и более 22 000 подтверждённых утечек в 145 странах; злоупотребление учётными данными осталось вторым по частоте вектором с долей 13 % [2]. Здравоохранение остаётся и самой дорогой отраслью: по отчёту IBM Cost of a Data Breach за 2026 г. средняя утечка в медицине обходится в 6,64 млн долл. против 4,99 млн долл. в среднем по миру, и это тринадцатый год подряд, когда отрасль занимает первое место по стоимости инцидента [3].

В США, где раскрытие утечек с числом пострадавших от 500 человек обязательно, в 2025 г. в Управление по гражданским правам Министерства здравоохранения поступило 710 сообщений о крупных утечках; защищённая медицинская информация как минимум 61 556 256 человек оказалась раскрыта, средний размер утечки составил 86 699 записей, медианный — 4 011 [4]. В Евросоюзе агентство ENISA разобрало 215 публично сообщённых инцидентов в здравоохранении: 54 % пришлось на программы-вымогатели, 42 % инцидентов затронули больницы, а данные пациентов, включая электронные медицинские карты, оказались самым востребованным активом — 30 % [5]. В обзоре ENISA Threat Landscape 2025 за период с июля 2024 по июнь 2025 г. на здравоохранение приходится 4,2 % выявленных киберпреступных инцидентов, при этом среди способов первичного проникновения доминирует фишинг, около 60 % наблюдавшихся случаев [6]. По Российской Федерации аналитический центр InfoWatch

поставил страну на второе место в мире по числу утечек из медицинских организаций в первом полугодии 2025 г. — 8,1 % всех зарегистрированных в мире инцидентов; доля внутренних нарушителей составила 21,4 % против 2,3 % в среднем по миру [7].

В Казахстане картина отличается качественно. Крупные утечки персональных данных граждан подтверждены официально: в июне 2025 г. в открытом доступе обнаружился массив из 16,3 млн записей, включавший 15,8 млн индивидуальных идентификационных номеров; Министерство цифрового развития, инноваций и аэрокосмической промышленности подтвердило подлинность данных, сообщив при этом, что фактов взлома государственных информационных систем не выявлено [8]. Одновременно цифровизация здравоохранения идёт быстро: в 2025 г. в электронном виде оказано 190,8 млн из 215,3 млн государственных медицинских услуг — 88 %; цифровая идентификация пациента внедрена в 1 950 организациях, в национальной цифровой системе зарегистрировано 275 тыс. медицинских работников [9]; к единому хранилищу данных подключено более двенадцати медицинских информационных систем, консолидируемых архитектурой e-Densaulyq [10]. Отраслевой статистики утечек медицинских данных при этом не публикуется. Само её отсутствие — результат: без обязательного раскрытия, сопоставимого с американским режимом, ни регулятор, ни отдельная клиника не могут оценить параметры, нужные для обоснования вложений в защиту, а обработка медицинских данных регулируется общими нормами Закона Республики Казахстан «О персональных данных и их защите» [11].

Во всех регионах, по которым есть данные, преобладающий путь проникновения проходит через учётные данные законного пользователя. Значит, релевантная мера — многофакторная аутентификация (МФА), и спорить следует не о том, внедрять ли её, а о том, сколько остаточного риска оставляет каждая схема и во что она обходится во времени врача.

2. Вероятностная модель

Принятая оценка остаточного риска схемы из n факторов трактует компрометацию отдельных факторов как независимые события:

$$P_{\text{нез}} = \prod q_i, \quad i = 1 \ldots n \tag{1}$$

где q_i — вероятность компрометации фактора i для целевой учётной записи в течение года.

На практике это не выполняется. Релей типа «противник посередине», работающий в реальном времени, за одно взаимодействие перехватывает пароль, одноразовый код и итоговый сессионный токен. Заражённый личный смартфон несёт одновременно менеджер паролей и приложение-аутентификатор. Подмена SIM-карты, удавшаяся в 39 попытках из 50 против пяти операторов связи по данным принстонского исследования — доля успеха 78 %, — разом обесценивает номер телефона и как второй фактор, и как канал восстановления пароля [12]. В терминах теории надёжности это отказы по общей причине; здесь они описаны бета-факторным разложением:

$$P_{\text{оп}} = (1 - \beta) \prod q_i + \beta\, q_1 \tag{2}$$

где β — доля атак на учётную запись, преодолевающих весь набор факторов через один канал, q_1 — вероятность компрометации первичного фактора. При $\beta = 0$ выражение (2) переходит в (1).

Величина β различается по схемам, причём различие структурное, а не эмпирическое. NIST SP 800-63B требует от аутентификатора, устойчивого к фишингу, чтобы его выход нельзя было использовать у подставной доверяющей стороны, даже если подписчика удалось склонить к аутентификации у неё; аутентификаторы на одноразовых паролях и внеполосные прямо названы неустойчивыми к фишингу, а внеполосная аутентификация через телефонную сеть общего пользования отнесена к ограниченным, и среди её факторов риска перечислены замена SIM-карты и перенос номера [13]. Утверждение FIDO2 криптографически привязано к источнику, поэтому канал релея не переносится: $\beta = 0$ здесь следует из построения, а не из допущения.

Сторона доступности описана последовательностью независимых попыток предъявления. Если p_i — вероятность ложного отказа фактора i за одну попытку, τ_i — среднее время его предъявления, то ожидаемое время одной успешной аутентификации при геометрическом повторении попыток равно

$$E[T] = \Sigma\, \tau_i / \prod (1 - p_i) \quad (3)$$

Для клиники с N учётными записями персонала вероятность компрометации хотя бы одной из них за год составляет

$$P_{клин} = 1 - (1 - P_{оп})^N \quad (4)$$

а полная ожидаемая годовая стоимость политики, складывающаяся из риска и потерянного врачебного времени,

$$C = P_{клин} \cdot L + N \cdot k \cdot E[T] \cdot c / 3600 \quad (5)$$

где L — средняя стоимость утечки, k — число аутентификаций на учётную запись в год, c — стоимость часа работы врача.

3. Параметры

Временные параметры взяты из измерений, а не из допущений. В исследовании, охватившем 19 больниц, 12 903 врача и 184 606 входов в систему за одну неделю, средний интерактивный вход по паролю занимал 34,6 с в начале смены и 29,3 с при повторном подключении [14]. Там же высвобожденное врачебное время оценено примерно в 113 долл. за час — величина получена из 10 302–13 245 часов стоимостью 1 164 126–1 496 685 долл. [14]. Частота входов k = 744 в год на учётную запись выведена из того же массива и намеренно занижена. Стоимость утечки L = 6,64 млн долл. — медицинский показатель из [3], профиль клиники N = 300 учётных записей. Вероятность ложного отказа, принятая для биометрического варианта, согласуется с требованием NIST: биометрическая система должна демонстрировать долю ложных несовпадений ниже 5 % [13]. Вероятности компрометации факторов и коэффициенты β сведены в таблицу 1; помеченные как принятые значения исследованы на чувствительность в разделе 6.

Таблица 1

Параметры модели

Фактор	q (1/год)	β	τ, с	p
Пароль	0,020	1,000	29,3	0,03
Одноразовый код по SMS	0,010	0,010	+20,0	0,05
Приложение TOTP	0,004	0,008	+10,0	0,02
Аппаратный ключ FIDO2	0,002	0,000	+3,0	0,01

4. Результаты и проверка модели

Сравниваются четыре схемы: S1 — только пароль; S2 — пароль и одноразовый код по SMS; S3 — пароль и приложение TOTP; S4 — пароль и аппаратный ключ FIDO2. В таблице 2 приведены остаточные вероятности компрометации на учётную запись в год по модели независимости (1) и по модели общей причины (2), их отношение, снижение относительно парольной базы и вероятность уровня клиники (4).

Таблица 2

Остаточная вероятность компрометации на учётную запись в год

Схема	P_нез	P_оп	P_оп / P_нез	Снижение	P_клин
S1 Только пароль	$2{,}00\cdot10^{-2}$	$2{,}00\cdot10^{-2}$	1,00	—	99,77 %
S2 + SMS	$2{,}00\cdot10^{-4}$	$3{,}98\cdot10^{-4}$	1,99	98,01 %	11,26 %
S3 + TOTP	$8{,}00\cdot10^{-5}$	$2{,}39\cdot10^{-4}$	2,99	98,80 %	6,93 %

Схема	P_нез	P_оп	P_оп / P_нез	Снижение	P_клин
S4 + FIDO2	$4{,}00 \cdot 10^{-5}$	$4{,}00 \cdot 10^{-5}$	1,00	99,80 %	1,19 %

Отсюда следуют два вывода. Для схем с одноразовыми паролями модель независимости занижает остаточный риск в два-три раза; для устойчивой к фишингу схемы обе модели совпадают, поскольку канал общей причины устранён по построению. Второй вывод существеннее: модель независимости оценивает преимущество S4 над S3 всего в 2,0 раза, тогда как модель общей причины даёт 5,98. Привычное допущение, стало быть, систематически недооценивает именно тот класс аутентификаторов, который рекомендует NIST.

Модель можно проверить по данным, не участвовавшим в её построении. Мейер с соавторами, разобрав коммерческий каталог учётных записей, где среди скомпрометированных 1 525 имели включённую МФА и 15 195 — отключённую, измерили снижение риска компрометации на 99,22 % по всей совокупности и на 98,6 % для записей с утёкшими паролями; там же доставка по SMS оказалась на 40,8 % менее эффективной, чем приложение-аутентификатор, при долях отказа 1,66 % против 0,97–0,99 % [15]. Модель раздела 2 даёт для схемы S3 снижение 98,80 %, попадающее внутрь измеренного интервала, и отношение остаточных рисков S2 к S3, равное 1,663, против измеренного отношения эффективности 1,69. Ни одна из этих величин не участвовала в подборе параметров, так что согласие служит независимой проверкой, а не круговым рассуждением.

Таблица 3 добавляет сторону доступности. Ожидаемое время аутентификации следует из (3), годовое врачебное время — из произведения $N \cdot k \cdot E[T]$, полная стоимость — из (5).

Таблица 3

Цена доступности и полная ожидаемая годовая стоимость

Схема	E[T], с	Часов/год	Риск, долл.	Время, долл.	Итого, долл.
S1 Только пароль	30,21	1 873	6 624 512	211 625	6 836 137
S2 + SMS	53,50	3 317	747 454	374 819	1 122 273
S3 + TOTP	41,34	2 563	460 142	289 644	749 786
S4 + FIDO2	33,64	2 085	79 205	235 649	314 854

Результат противоречит расхожему представлению, будто усиление аутентификации непременно утяжеляет работу. Схема S2 доминируется по Парето: среди схем с МФА она оставляет наибольший остаточный риск и одновременно съедает больше всего врачебного времени — около 3 317 часов в год на клинику в 300 учётных записей, поскольку в ожидаемое время входят и ожидание сообщения, и сбои его доставки. Схема S4 минимизирует обе составляющие сразу и снижает полную ожидаемую годовую стоимость в 2,4 раза относительно S3 и в 3,6 раза относительно S2. Для клиники устойчивый к фишингу вариант оказывается дешёвым, а не дорогим.

5. Адаптивная по риску политика

Адаптивную по риску аутентификацию часто предлагают именно для клиник — она обещает снять трение с рутинных сессий. В распространённом варианте политики сильный фактор требуется лишь тогда, когда механизм оценки риска счёл сессию подозрительной, а в остальных случаях хватает пароля. Если d — вероятность того, что механизм верно классифицирует сессию нарушителя, остаточный риск политики равен

$$P_{ад} = d \cdot P_{сильн} + (1 - d) \cdot P_{слаб} \qquad (6)$$

поскольку противник идёт по тому пути, который механизм оставил открытым. В таблице 4 выражение (6) посчитано при S4 в сильной ветви и одном пароле в качестве отката.

Таблица 4

Адаптивная политика с откатом на один пароль

d	P_ад	P_клин	к S3	к S4
0,800	$4{,}03\cdot10^{-3}$	70,24 %	в 16,8 раза хуже	в 100,8 раза хуже
0,900	$2{,}04\cdot10^{-3}$	45,74 %	в 8,5 раза хуже	в 50,9 раза хуже
0,950	$1{,}04\cdot10^{-3}$	26,77 %	в 4,3 раза хуже	в 26,0 раза хуже
0,990	$2{,}40\cdot10^{-4}$	6,94 %	наравне	в 6,0 раза хуже
0,999	$6{,}00\cdot10^{-5}$	1,78 %	в 4 раза лучше	в 1,5 раза хуже

Решение (6) относительно точки безразличия даёт $d \geq 99{,}00$ %, чтобы адаптивная политика лишь сравнялась с равномерной S3; чтобы сравняться с равномерной S4, потребовалось бы $d = 100$ %, что недостижимо. Механизм такой точности в клинической сети нереалистичен. Законные шаблоны доступа там нерегулярны по своей природе: ночные смены, совместители, общие рабочие станции, экстренный доступ — всё это внешне неотличимо от поведения нарушителя. Практический вывод: адаптация должна работать между устойчивыми к фишингу методами, меняя, какой именно сильный аутентификатор требуется или как долго живёт сессия, и никогда — между многофакторной и однофакторной аутентификацией. То же рассуждение относится к процедурам экстренного доступа «разбить стекло», которые снимают требование факторов целиком; компенсировать их следует обязательной отчётностью постфактум, а не ослаблением пути аутентификации [16].

6. Чувствительность

Коэффициент β — наименее наблюдаемый параметр модели, поэтому в таблице 5 отношение P_оп / P_нез посчитано на диапазоне его значений. Качественный вывод устойчив: при любом β выше 0,002 модель независимости занижает остаточный риск схем с одноразовыми паролями, и занижение растёт вместе с силой второго фактора. В этом и состоит контринтуитивная суть результата: чем лучше второй фактор сам по себе, тем большая доля остающегося риска сосредоточена в канале общей причины и тем сильнее вводит в заблуждение допущение независимости.

Таблица 5

Коэффициент занижения P_оп / P_нез как функция β

β	S2 (SMS)	S3 (TOTP)	S4 (FIDO2)
0,000	1,00	1,00	1,00
0,002	1,20	1,50	2,00
0,005	1,49	2,25	3,50
0,010	1,99	3,49	5,99
0,020	2,98	5,98	10,98
0,050	5,95	13,45	25,95

Изменение базовой вероятности компрометации пароля q_1 в диапазоне 0,005–0,050 меняет абсолютные цифры, но не порядок схем: вероятность уровня клиники для S4 остаётся между 0,30 % и 2,96 %, тогда как для S1 она нигде не опускается ниже 77,8 %. Ранжировка $S4 \prec S3 \prec S2 \prec S1$ сохраняется на всём исследованном диапазоне.

7. Выводы

Считать факторы схемы аутентификации независимо компрометируемыми удобно, но неверно, причём ошибка смещена в опасную сторону. Для схем с одноразовыми паролями модель независимости занижает остаточный риск в два-три раза и втрое недооценивает преимущество аутентификаторов, устойчивых к фишингу. Предложенное бета-факторное описание воспроизводит две независимо измеренные величины — 98,80 % расчётного снижения против 98,6–99,22 % измеренных и расчётное отношение 1,663 между схемами на SMS и на приложении против измеренного 1,69, — не будучи подогнанным ни под одну из них.

Для клиники в 300 учётных записей практических выводов три. Аутентификацию одноразовым кодом по SMS следует вывести из эксплуатации: она проигрывает сразу по обеим осям. Устойчивые к фишингу аутентификаторы снижают полную ожидаемую годовую стоимость в 2,4 раза относительно приложения-аутентификатора и в 3,6 раза относительно SMS, так что они оправданы чисто экономически, без апелляции к требованиям регуляторов. Адаптивные политики не должны откатываться к одному фактору: требуемая точность механизма оценки риска, не ниже 99 %, превышает то, что позволяет клинический шаблон доступа.

Для Казахстана отдельный вывод следует из того, что посчитать не удалось. Все параметры рисковой стороны модели пришлось взять из американских и европейских источников, потому что отраслевой статистики утечек медицинских данных в стране не публикуют — при том что архитектура e-Densaulyq консолидирует медицинские информационные системы почти двух тысяч организаций. Обязательный режим раскрытия утечек медицинских данных, сопоставимый по принципу с порталом Управления по гражданским правам США, — предпосылка доказательных решений по защите в национальном здравоохранении и направление дальнейшей работы автора.

ЛИТЕРАТУРА

1. *Verizon Business.* Healthcare snapshot. 2026 Data Breach Investigations Report. – Basking Ridge, NJ: Verizon, 2026. – https://www.verizon.com/business/resources/reports/2026-dbir-healthcare-snapshot.pdf (дата обращения: 15.09.2026).

2. *Verizon Business.* Executive Summary. 2026 Data Breach Investigations Report. – Basking Ridge, NJ: Verizon, 2026. – https://www.verizon.com/business/resources/executivebriefs/2026-dbir-executive-summary.pdf (дата обращения: 15.09.2026).

3. *IBM Security.* Cost of a Data Breach Report 2026. – Armonk, NY: IBM Corporation, 2026. – Цит. по: Global data breach cost rises 12 % to almost $5 million // The HIPAA Journal. – https://www.hipaajournal.com/2026-cost-data-breach-study-ibm/ (дата обращения: 15.09.2026).

4. *Alder S.* 2025 Healthcare Data Breach Report // The HIPAA Journal. – 2026. – По данным Управления по гражданским правам Министерства здравоохранения и социальных служб США. – https://www.hipaajournal.com/2025-healthcare-data-breach-report/ (дата обращения: 15.09.2026).

5. *ENISA.* Health Threat Landscape. – Афины: Агентство Европейского союза по кибербезопасности, 5 июля 2023. – https://www.enisa.europa.eu/news/checking-up-on-health-ransomware-accounts-for-54-of-cybersecurity-threats (дата обращения: 15.09.2026).

6. *ENISA.* ENISA Threat Landscape 2025. – Афины: Агентство Европейского союза по кибербезопасности, октябрь 2025. – https://www.enisa.europa.eu/sites/default/files/2026-01/ENISA%20Threat%20Landscape%202025_v1.2.pdf (дата обращения: 15.09.2026).

7. *Экспертно-аналитический центр InfoWatch.* Утечки конфиденциальных данных из медицинских организаций. Мир – Россия, I полугодие 2025 г. – М.: InfoWatch, 2025. – https://habr.com/ru/companies/infowatch/news/937060/ (дата обращения: 15.09.2026).

8. Киберугроза: почему произошла утечка персональных данных казахстанцев и как их защитить // Kazinform. – Астана, 1 августа 2025. – https://www.inform.kz/ru/kiberugroza-pochemu-proizoshla-utechka-personalnih-dannih-kazahstantsev-i-kak-ih-zashitit-df15ec (дата обращения: 15.09.2026).

9. *Министерство здравоохранения Республики Казахстан.* В Казахстане 88 % госуслуг в сфере медицины переведены в электронный формат. – Астана, 21 августа 2026. – https://ru.sputnik.kz/20260821/v-kazakhstane-88-gosuslug-v-sfere-meditsiny-perevedeny-v-elektronnyy-format-65891326.html (дата обращения: 15.09.2026).

10. *Profit.kz.* В Казахстане усиливают цифровизацию системы здравоохранения. – Алматы, 12 мая 2026. – https://profit.kz/news/73715/V-Kazahstane-usilivaut-cifrovizaciu-sistemi-zdravoohraneniya/ (дата обращения: 15.09.2026).

11. Закон Республики Казахстан от 21 мая 2013 года № 94-V «О персональных данных и их защите» (с изменениями и дополнениями). – Астана, 2013. – https://adilet.zan.kz/rus/docs/Z1300000094 (дата обращения: 15.09.2026).

12. *Lee K., Kaiser B., Mayer J., Narayanan A.* An Empirical Study of Wireless Carrier Authentication for SIM Swaps // Proceedings of the Sixteenth USENIX Symposium on Usable Privacy and Security (SOUPS 2020). – Berkeley, CA: USENIX Association, 2020. – P. 61–79.

13. *Temoshok D., Fenton J., Choong Y.-Y., Lefkovitz N., Regenscheid A., Galluzzo R., Richer J.* Digital Identity Guidelines: Authentication and Authenticator Management. NIST Special Publication 800-63B-4. – Gaithersburg, MD: National Institute of Standards and Technology, 31 июля 2025. – DOI: 10.6028/NIST.SP.800-63b-4. – https://pages.nist.gov/800-63-4/sp800-63b.html (дата обращения: 15.09.2026).

14. *Gellert G.A., Delacerda C., Patel L., Maciaz G.* Easing hospitalist electronic health record burden through clinical workstation single sign-on // Journal of Hospital Administration. – 2020. – Vol. 9, No. 3. – P. 24. – DOI: 10.5430/jha.v9n3p24.

15. *Meyer L.A., Romero S., Bertoli G., Burt T., Weinert A., Lavista Ferres J.* How effective is multifactor authentication at deterring cyberattacks? – Redmond, WA: Microsoft Research, 2022. – https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/MFA-Microsoft-Research-Paper-update.pdf (дата обращения: 15.09.2026).

16. *Al Amin M., Shah R., Tummala H., Ray I.* Did You Break the Glass Properly? A Policy Compliance Framework for Protected Health Information Emergency Access // Proceedings of the 22nd International Conference on Security and Cryptography (SECRYPT 2025). – Setúbal: SciTePress, 2025.

Federated Learning for Privacy-Preserving ESG Risk Assessment in Cloud-Native Financial Systems

Suresh Varma Dendukuri
Anna University, India
sureshvashishtad@gmail.com

Abstract - The distributed and sensitive nature of ESG data poses privacy and regulatory challenges, while also serving as a critical piece of financial risk evaluation. This paper introduces a privacy-preserving federated learning-based collaborative ESG risk assessment framework in a cloud-based financial system. The framework combines Federated Learning, Differential Privacy, and Secure Aggregation that allows multiple financial institutions to jointly create ESG risk models without sharing raw data. A scalable and efficient model management is enabled by incorporating a cloud-native deployment architecture, built on containerized services and Kubernetes orchestration. Through the experimental study based on the ESG-related dataset, the proposed framework is evaluated and the accuracy is shown as 89.3% after 50 communication rounds, whereas the accuracy of the conventional federated learning and centralized learning are 90.0% and 91.8%, respectively. The framework offers greater privacy and secure collaborative learning without sacrificing prediction accuracy, but with a slight drop in accuracy compared to the previous model. The results show that it is possible to create privacy-preserving ESG intelligence for distributed financial systems.

Keywords- Federated Learning, Environmental Social and Governance (ESG), Privacy-Preserving Analytics, Differential Privacy, Cloud-Native Financial Systems.

I. INTRODUCTION

Environmental, Social, and Governance (ESG) assessment has become a critical tool in the strategic risk evaluation process, not just a reporting activity as it was previously. Businesses and financial institutions are increasingly using ESG metrics to gauge organisational resilience, look for long-term sustainability risks and to make responsible investment decisions. With the increased importance of ESG considerations for lenders, portfolio management and regulatory requirements, reliable and complete ESG risk assessments are becoming more critical than ever. The ability to obtain meaningful ESG evaluation often requires information across various financial institutions such as banks, investment companies, insurance companies and regulatory bodies. These entities have a variety of environmental performance data, as well as data on workforce practices, governance, compliance and financial disclosures. While this can enhance the quality and reliability of ESG assessments, direct data sharing is often limited by privacy regulations, institutional policies, and competitiveness concerns. Many current ESG evaluation methods, therefore, have been based on partial or incomplete data sets, which are unable to provide a full picture of sustainability-related risks.

A promising alternative is Federated Learning (FL), which allows collaborative development of a model without the exchange of raw data among participating institutions. Orgs train models locally and only send updates once to model for global aggregation, rather than centralizing sensitive info. This decentralized learning approach opens up a new potential for large-scale ESG risk assessment without compromising data privacy and enabling regulatory compliance. While ESG analytics and privacy preserving machine learning methods continue to be used, FL for ESG risk assessment has not been explored to the same extent. Moreover, in today's financial landscape, distributed and cloud-native infrastructures are essential to support the training of distributed models, secure communication, and model updates across geographically dispersed organizations. Current approaches are not often able to tackle these requirements under a unified framework that protects privacy, leaving there a major research gap on the intersection of ESG analytics, federated learning, and cloud-native computing. In

response, the authors presented a Federated Learning framework for privacy-preserving ESG risk assessment in cloud-native financial systems in this paper. The proposed architecture allows several financial institutions to build a joint financial ESG risk model without sharing any sensitive local data. To enhance data security, especially during model training and aggregation, differential privacy and secure aggregation techniques are integrated, and a cloud-native architecture ensures scalable deployment and efficient model management in distributed environments.

This work has four key contributions. First, a federated learning framework is put forward for cooperative ESG risk assessment among different financial institutions without sharing raw data so that the data of each financial institution can be retained and the information can be kept confidential. Secondly, privacy-preserving techniques such as differential privacy and secure aggregation are embedded in the learning process to protect sensitive information and prevent potential privacy leakage in model training and communication. Third, a cloud-native deployment architecture has to be scalable, secure, and efficient to enable federated learning operations in financial distributed context via containerized services and orchestration mechanisms. Finally, the effectiveness of the proposed framework is assessed by extensive experiments, covering prediction performance, privacy preservation, model convergence behavior and communication overhead, proving its applicability for private ESG risk assessment in today's financial world.

II. RELATED WORK

In recent years, the growing significance of Environmental, Social, and Governance (ESG) assessment has prompted researchers to investigate how data can be leveraged for sustainable financial decision-making. Lim [1] gave a detailed review of the application of artificial intelligence in the financial sector for ESG and emphasized the increasing importance of machine learning techniques in the context of ESG assessment, in terms of both improving the accuracy and scalability of such assessments. Likewise, Martin-Melero et al. [15] explored the application of machine learning to ESG scores of the mutual funds, and showed that integration of such sector-specific and financial indicators can improve the performance of ESG scores. These studies could be classified as the early beginnings of ESG analytics but were largely built around centralized data processing architectures that needed to pull and consolidate data from various sources.

Recognizing the privacy concerns in centralized financial analytics, Federated Learning (FL) has emerged as a promising paradigm for collaborative model training without sacrificing privacy. In financial applications, the feasibility of distributed model training is illustrated by dasari and kaluri [11] who proposed a privacy-preserving framework that leverages Flower Federated Learning. A federated credit-scoring approach was suggested by Wang et al. [12] with the introduction of knowledge transfer mechanisms which enhance the predictive performance of the credit scoring system by distributed institutions. Moreover, Jiang et al. [14] proposed the FedRisk system for financial risk assessment on cloud platforms, demonstrating the capability of collaborative learning to enhance financial risk prediction without compromising data confidentiality among financial institutions. Gajula extended financial analytics even further to introduce a federated intelligence system for financial risk analysis across borders, which highlights the need for collaboration and security among geographically dispersed entities.Gajula [13] introduced a federated intelligence system for financial risk analysis across borders, which extends the scope of financial analytics, giving emphasis to collaboration and security in the financial risk analysis among geographically distributed entities.

While FL has its benefits, privacy leakage via model updates is a major concern. In order to better protect privacy, some studies have introduced differential privacy and its related techniques into federated learning systems. Naresh and Ayyappa [2] tested differential privacy and homomorphic encryption methods in federated credit risk models and showed their potentials in minimizing information leakage. To address the trade-off between privacy protection and accuracy, Feng et al. [3] proposed PPRD-FL, a privacy-preserving federated

learning framework based on randomized parameter selection and dynamic local differential privacy, which can improve privacy protection while ensuring high accuracy. Yin et al To achieve even better privacy preservation, [4] proposed a dual-mode switching mechanism that adjusts the differential privacy level as needed to balance privacy and model utility in light of data drift. Similarly, Yang [5] investigated the optimization strategies for differential privacy in the context of financial risk control, and he found that one could achieve better privacy guarantees while having little degradation in predictive accuracy.

Another research direction of great importance is security of aggregation in federated environments. To safeguard the model updates from unauthorized access and manipulation in collaborative learning, Kanchan et al. [6] proposed a group signature-based secure aggregation framework (GSfedSec). To enhance the robustness and trustworthiness of distributed learning environments, Sun [9] proposed a federated financial risk control security-weighted aggregation strategy. Furthermore, Ma et al. [10] proposed a federated supply-chain finance risk control framework with personalized differential privacy, which highlighted the potential of applying privacy-preserving learning methods in complex financial scenarios. There are also recent studies that focused on how to make federated systems more efficient, scalable, etc. In order to minimize the communication overhead and boost the efficiency of training, Guo [7] presented a financial data management framework based on data distillation and dynamic tiered scheduling. Advanced federated architectures could be applied to real-world financial services, as seen in Zheng's [8] creative fusion of adaptive federated learning and generative adversarial networks for fraud detection and data protection in banking systems.

Current work in ESG analytics, federated financial learning, differential privacy, and secure aggregation has achieved major strides in each of them, but they have been studied in isolation. The majority of the current ESG assessment frameworks are based on the centralized approach to data aggregation, whereas the federated financial risk studies have been primarily limited to credit scoring, fraud detection, and risk management. Based on our observations, there have been only a few studies that have examined collaborative ESG risk assessment in a federated learning framework, combined with privacy-preserving solutions, and deployed on the cloud. This is the motivation behind the proposed privacy-preserving federated system for ESG risk assessment in cloud-based financial systems.

III. PROPOSED SYSTEM ARCHITECTURE

The framework is proposed to allow for the ESG risk assessment process to take place across multiple financial institutions in a collaborative manner with no sharing of sensitive organizational information. The proposed approach is a decentralized learning paradigm in which every participating institution maintains full control over its local data, thereby providing their contribution in the development of a common global intelligence model, rather than relying on centralized data repositories, as traditional ESG analytics platforms do. The design not only meets privacy and regulatory requirements but also enhances the scalability and usability of ESG evaluation in decentralized financial systems.

The entire architecture of the proposed framework is shown in Figure 1. It includes the financial institutions that are part of the system, local ESG model training environments, privacy-preserving communication modules, a federated aggregation server, cloud-native orchestration services, and a global ESG intelligence layer. All individual institutions hold their own ESG-related data such as environmental data, social responsibility data, governance data, compliance data, financial disclosures, etc. Raw data do not cross institutional boundaries, meaning data ownership is maintained and collaborative learning is still possible.For each institution, ESG features are extracted and used to train local machine learning models. The models learn patterns of sustainability-related risk from the institution-specific datasets and produce model parameters that are representative of local knowledge. Only model updates are created for communication with the federated aggregation server, rather than sending sensitive records. Such a mechanism can substantially mitigate privacy concerns linked to centralized data

collection, and let institutions reap the benefits of collective intelligence. Furthermore, model updates are processed through a privacy preserving layer prior to being transmitted, to further enhance confidentiality. Differential privacy mechanisms add carefully controlled noise to the parameters of the model, and secure aggregation protocols involve adding noise to individual contributions when aggregating different contributions.

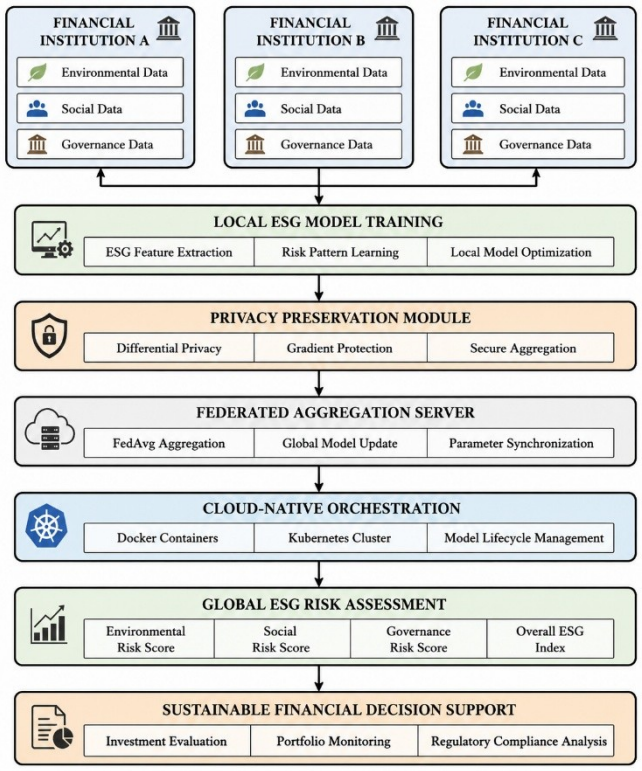


Fig. 1. Proposed Privacy-Preserving Federated ESG Risk Assessment Framework for Cloud-Native Financial Systems.

Thus, the aggregation server does not need to access the raw data or institution-specific model information to build a global model. The federated aggregation server serves as the "hub" of the aggregating framework. The server receives model updates from the participating institutions after providing them with protection, and then it merges the models to create an enhanced global ESG model. The aggregation process is carried out following the Federated Averaging (FedAvg) strategy, which averages local model parameters based on the size of local training datasets. The global update to the model can be represented by

$$W^{t+1} = \sum_{k=1}^{N} \frac{n_k}{n} W_k^t \quad (1)$$

In the above, the local model parameters of the kth participating institution in the t-th communication round are represented by W_k^t, the number of local training samples at the kth participating institution is n_k, and the total number of participating institutions is N. This weighted aggregation approach allows for sharing of information, but does not allow for any raw data to be shared. This leads to a global model which is then returned to the participating institutions for the next training cycle. The global ESG intelligence model is developed across several rounds of communication until the criterion of convergence is met. Through this ongoing learning cycle, institutions with different ESG datasets can collaborate and collectively enhance prediction capabilities without compromising on data privacy or regulatory adherence.

The framework also features cloud-native orchestration services for model lifecycle management, deployment automation, and resource allocation, enabling seamless and scalable operations. The framework also integrates with cloud-native orchestration services for model lifecycle management, deployment automation, and resource allocation, ensuring seamless and scalable operations. Containerized federated learning services enable easy integration between geographically distributed institutions, and orchestration services ensure a continuous update of

the models and efficient coordination between participating entities. Finally, the model created for the whole world is used to analyze environmental performance, social responsibility and governance compliance, creating all in one ESG risk scores that will be able to assist in investment analysis, portfolio monitoring, regulatory evaluation and sustainability-focused investment decision.

IV. PRIVACY-PRESERVING ESG RISK ASSESSMENT

It proposes a framework that integrates two techniques Differential Privacy (DP) and Secure Aggregation (SA) to allow collaborative ESG risk assessment, while preserving sensitive institutional information. These mechanisms provide for participation without the sharing of confidential data in model development.

A. Differential Privacy Mechanism

While Federated Learning does not share raw data, it can still provide updates to models that may contain sensitive data. To reduce this hazard, it is possible to apply Differential Privacy to local model parameters before their transmission. Noise is added to the model updates to produce controlled Gaussian noise.

$$\widetilde{W}_k = W_k + \mathcal{N}(0, \sigma^2) \tag{2}$$

In this, $\widetilde{W}_k$ denotes the local model update with privacy protection, W_k represents the original local model parameters, and $\mathcal{N}(0, \sigma^2)$denotes Gaussian noise with variance σ^2. This additional noise makes it more difficult to deduce the information that is sensitive while retaining the value of the learning process.

B. Secure Aggregation Strategy

Secure Aggregation also is used during model communication to further improve privacy. Protected model updates from each institution are masked/encrypted prior to transmission. As a result, the aggregation server is only able to retrieve the aggregated result and not each individual contribution. This helps to avoid the possibility of unauthorized reconstruction of local information and promotes confidence of the participating institutions.

C. ESG Risk Scoring Model

The global model is used after federated training to calculate ESG scores based on environmental, social and governance metrics. The overall ESG score is calculated as:

$$ESG = w_E E + w_S S + w_G G \tag{3}$$

This is because it involves the Environmental E, Social S and Governance G scores and their weights(w_E, w_S and $w_G$respectively). The higher ESG is, the more sustainable the company is and the less it is exposed to overall risk.

Table 1 provides an overview of the ESG dimensions, their representative indicators and data sources in the proposed framework.

Figure 2 shows the proposed federated ESG assessment framework deployment architecture in the cloud. The deployment environment is aimed to make model training scalable, secure and automated, and to create a continuous model management in geographically distributed financial institutions. Cloud-native technologies facilitate coordination of containerized federated learning services for efficient resource usage and collaboration among the entities participating. The proposed framework is implemented in a cloud-native approach to ensure scalability, flexibility and operational efficiency in distributed financial environments. The participating institutions have a federated learning client running that is containerized and provides local model training and secure communication with an aggregation server. By leveraging containerization, applications can run on any platform, and updates and maintenance of models are made more convenient.

TABLE 1. ESG SCORING DIMENSIONS AND DATA SOURCES

ESG Dimension	***Representative Indicators***	***Data Sources***

Environmental	Carbon Emissions, Energy Consumption, Waste Management	Sustainability Reports, Environmental Audits
Social	Employee Welfare, Workforce Diversity, Community Engagement	HR Records, CSR Reports
Governance	Regulatory Compliance, Board Structure, Ethical Policies	Compliance Reports, Corporate Governance Records

V. CLOUD-NATIVE IMPLEMENTATION

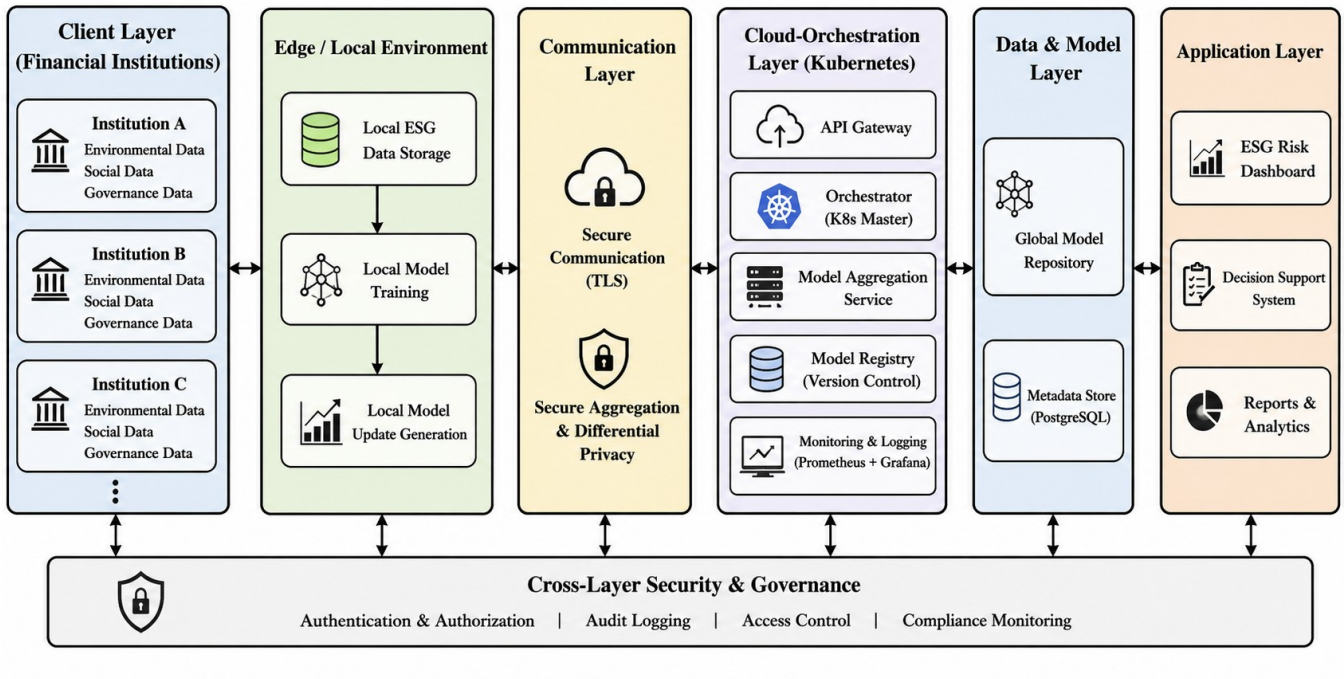


Fig. 2. Cloud-Native Deployment Architecture for Federated ESG Risk Assessment.

A Kubernetes-based orchestration layer takes care of the lifecycle of federated learning services: resource allocation, workload balancing, service discovery, and automatic scaling. This orchestration mechanism guarantees the reliability of operation in case several institutions are involved in the federated learning process at the same time. Moreover, the containerized services enable rapid deployment and seamless integration to heterogeneous computing environments. To enable efficient communication, the model also uses event-driven messaging services to ensure that the aggregation server and the federated client are synchronized about changes to the model. Centralized model registry supports versioning and lifecycle management of models to ensure access to the most current validated global model by institutions. Additionally, Continuous Integration and Continuous Deployment (CI/CD) pipelines streamline the deployment and update process of models, further automating operational procedures and enhancing system responsiveness. The cloud-native design allows the proposed framework to scale to support more institutions, and to ensure that performance, security, and availability remain stable. This, in turn, enables an efficient and usable system to

evaluate ESG risks on a large scale in distributed financial ecosystems while maintaining privacy.

VI. Results and Discussion

A. Dataset and Experimental Setup

A set of environmental, social, governance, and financial indicators from publicly listed companies was used to assess the proposed framework using the publicly available ESG dataset. The data covers a range of carbon emissions, energy use, workforce diversity, employee health and safety, regulatory compliance and governance.

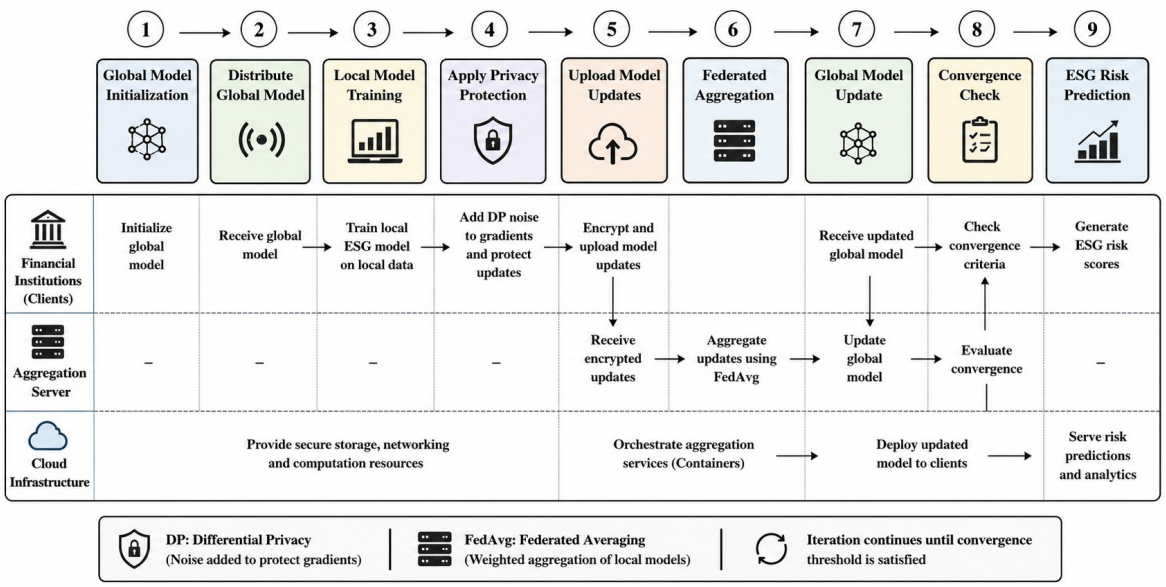


Fig. 3. Workflow of the Privacy-Preserving Federated ESG Risk Assessment Process.

In order to create a financial ecosystem that is more collaborative, the dataset was shared among five of the participating institutions, each of which had its own local data repository. The data was preprocessed and normalized before training, and then split into an 80–20 train–test ratio. Local ESG models were trained for 5 epochs in each communication round and the Federated Averaging (FedAvg) algorithm was used for aggregating the global model across 50 communication rounds. Differential Privacy and Secure Aggregation were incorporated to ensure privacy throughout the learning process. The entire framework was implemented in a cloud-native environment based on Kubernetes, and the framework was considered by prediction accuracy, convergence behavior, privacy preservation and communication overhead.

B. Workflow Execution and Model Convergence

Figure 3 shows the operational workflow of the proposed federated ESG assessment framework, which is based on privacy. The operational workflow of the proposed federated ESG assessment framework based on privacy is illustrated in Figure 3. This process starts with the global model initialization followed by local ESG model training at the participating institutions. Privacy-preserving mechanisms are used prior to sending model updates to the aggregation server. This consolidated global model is then multiplied and passed to the participating institutions and the process is repeated until convergence occurs. The workflow represents one after another local training, privacy protection, secure aggregation and global model refinement. The framework gradually enhances the prediction accuracy with an increasing number of communication rounds while retaining data privacy. Experimentally, the convergence behavior shows the successful performance of the federated learning procedure in creating trustworthy ESG risk estimates in distributed financial environments.

C. Privacy and Performance Analysis

A privacy–utility trade-off analysis was performed to see how well the model performs at varying privacy budgets to assess the impact of preserving privacy on the model performance.

The relationship between prediction accuracy and privacy protection with varying privacy budget (ε) is shown in Figure 4. The results show that higher privacy guarantees are achieved at lower privacy budgets, and higher privacy budgets at increased prediction accuracy. While recognizing the compromise, the suggested strategy keeps its ESG prediction performance competitive and offers strong privacy security, making it suitable for privacy-sensitive monetary systems.

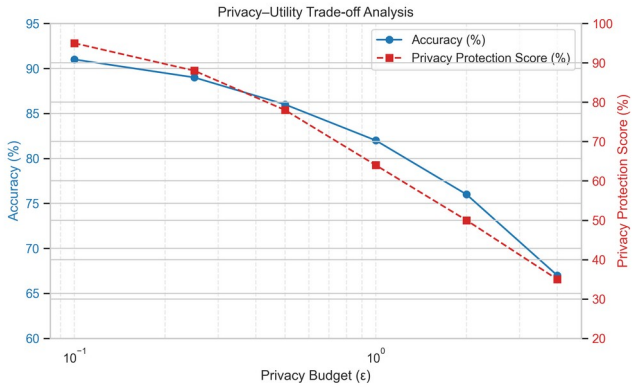


Fig. 4. Privacy–Utility Trade-off Analysis under Different Privacy Budget Settings

As shown in Figure 4, prediction accuracy improves smoothly as privacy budgets grow, as the updates to the models are less disturbed. On the other hand, the more noise that is removed from the learning process, the less privacy is protected. The results validate the proposed framework which allows for the preservation of privacy while safeguarding the predictive utility of the model, while ensuring that the model performance remains unaffected.

D. ESG Component Analysis

To explore the sustainability aspects that can by captured by the proposed framework an institution-wise analysis of Environmental, Social and Governance (ESG) scores was performed. Figure 5 shows the breakdown of the ESG dimensions among the institutions participating. The analysis also enables insights into the variability of sustainability indicators and the different ESG profiles in the collaborative financial ecosystem.

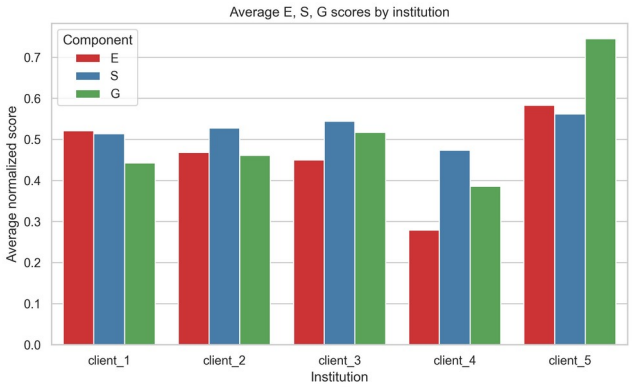


Fig. 5. Institution-Wise Distribution of Environmental, Social, and Governance Scores.

Figure 5 shows that there are significant differences among the institutions regarding environmental, social and governance aspects. Some institutions are showing good results in environmental performance, whereas others are showing comparatively good results in social or governance performance. These differences highlight the diversity of the information available on the ESG issue and the need for joint education for producing a holistic ESG risk assessment. The proposed framework combines knowledge from various institutions, which enables it to cover various sustainability characteristics and create a more well-rounded ESG evaluation framework.

D. Comparative Performance Evaluation

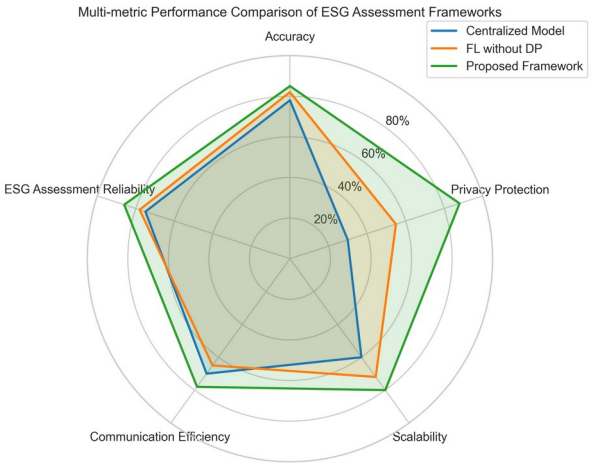


Fig. 6. Multi-Metric Performance Comparison of the ESG Assessment Frameworks.

In order to capture a full view of the proposed framework, several performance dimensions were considered other than prediction accuracy. Fig. 6 shows comparative analysis of the proposed framework with conventional centralized and federated learning (without privacy protection). The comparison focuses on the most important aspects such as prediction accuracy, privacy protection, scalability, communication efficiency, and the reliability of the ESG assessment, all of which are essential to the practical applicability of ESG risk assessment systems in a distributed financial system. The proposed framework is capable of achieving the balance between all the evaluation metrics as shown in Figure 6. It provides significant enhancements in privacy protection, scalability, and reliability of ESG assessment over the baseline methods without compromising prediction accuracy. The use of secure aggregation and cloud-native orchestration in the framework helps to achieve secure collaborative learning while maintaining operational efficiency, making it applicable to complex, large-scale ESG risk assessment for multiple financial institutions.

E. Overall Performance Comparison

The overall performance of the proposed framework is summarized in Table 2, and compared with centralized learning and conventional federated learning approaches. The comparison takes into account the accuracy of predictions, privacy settings, and communication overhead. The centralized model obtains a slightly better accuracy, but it requires full data centralisation and has high communication costs. The proposed framework preserves privacy, scalable distributed learning capability and competitive prediction accuracy.

TABLE 1. PERFORMANCE COMPARISON OF ESG ASSESSMENT APPROACHES

Method	***Accuracy (%)***	***Privacy Budget (ε)***	***Communication Cost (MB)***
Centralized ESG Model	91.8	N/A	850
FL without Privacy Protection	90.0	N/A	420
Proposed Framework	89.3	1.0	465

The results show that the proposed framework provides a good trade-off between privacy control, predictive accuracy and communication efficiency. A slight decrease in accuracy is noted due to the incorporation of privacy-preserving techniques, but the framework still manages to remove the necessity for the raw data to be shared and enables safe multi-

institutional collaborative ESG assessments. The decreased amount of communication overheads also indicates its relevance for cloud-native financial settings.

VII. CONCLUSION AND FUTURE ENHANCEMENTS

This paper proposed a federated learning framework that preserves privacy for the ESG risk assessment within cloud based financial system. The proposed solution allows for multiple financial institutions to create ESG intelligence without sharing sensitive organizational data, solving privacy, regulatory compliance, and data sharing problems in the typical ESG analytics process. The framework incorporates Differential Privacy and Secure Aggregation into the federated learning framework, safeguarding confidential information and maintaining the collaborative model development benefits. Moreover, the cloud-native deployment architecture offers scalable and flexible distributed training, model administration and resource orchestration. The proposed framework proved to be effective in terms of its performance in terms of predicting the local and global maximum and in preserving privacy as shown by experimental results. The proposed model had an accuracy of 89.3% after 50 rounds of communication, whereas conventional federated learning and centralized learning were able to achieve an accuracy of 90% and 91.8% respectively. While there was some loss of accuracy as a result of the privacy-preserving mechanisms, the framework was able to maintain high levels of confidentiality without requiring raw data to be exchanged. The privacy–utility analysis also clearly showed that secure collaborative learning is possible without seriously compromising the accuracy of the model. Furthermore, the comparative assessment has shown that the process is scalable, efficient in terms of communication, and showed that the process is reliable in assessing the ESG of a financial ecosystem.

This could be followed by adaptive ESG weighting strategies, which can adapt to changing regulatory and market conditions, with the contribution of each indicator of ESG varied dynamically. Other research should explore communication-efficient federated optimization methods, empowering support for highly heterogeneous institutional datasets, and cross-border ESG assessment across a variety of compliance frameworks. Moreover, the incorporation of EAI mechanisms can help increase transparency and increase the depth of understanding on ESG risk prediction, building trust and better decision-making in finance for sustainable systems.

REFERENCES

[1] Lim, Tristan. "Environmental, social, and governance (ESG) and artificial intelligence in finance: State-of-the-art and research takeaways." Artificial Intelligence Review 57, no. 4 (2024): 76.

[2] Naresh, Vankamamidi S., and D. Ayyappa. "Privacy-preserving federated credit risk models: evaluating differential privacy and homomorphic encryption techniques." Scientific Reports (2026).

[3] Feng, Jianlong, Rongxin Guo, and Jianqing Zhu. "PPRD-FL: Privacy-Preserving Federated Learning Based on Randomized Parameter Selection and Dynamic Local Differential Privacy." Electronics 14, no. 5 (2025): 990.

[4] Yin, Chunxiao, Kai He, Zhiyang Li, Ninghang Kang, and Jiaoli Shi. "Leveraging data drift to optimize differential privacy in federated learning via dual-mode switching." Journal of King Saud University Computer and Information Sciences 37, no. 8 (2025): 225.

[5] Yang, Ziang. "Privacy-Aware Financial Risk Control: A Federated Learning Approach with Differential Privacy Optimization." Journal of Computer Technology and Software 4, no. 4 (2025).

[6] Kanchan, Sneha, Jae Won Jang, Jun Yong Yoon, and Bong Jun Choi. "GSFedSec: Group signature-based secure aggregation for privacy preservation in federated learning." Applied Sciences 14, no. 17 (2024): 7993.

[7] Guo, Shiqin. "Federated learning financial data management techniques based on data distillation and dynamic tiered scheduling." Discover Applied Sciences 7, no. 12 (2025): 1406.
[8] Zheng, Yunjiao. "Bank data protection and fraud identification based on improved adaptive federated learning and wgan." Scientific Reports 15, no. 1 (2025): 23006.
[9] Sun, Xiaowanqing. "Research on the Application of Security-Weighted Aggregation in Federated Learning for Financial Risk Control." Finance & Economics 1, no. 6 (2025).
[10] Ma, C., H. Zhao, K. Zhang, L. Zhang, and H. Huang. A federated supply chain finance risk control method based on personalized differential privacy. Egyptian Informatics Journal 31 (1), 100704 (2025). 2025.
[11] Dasari, Sandeep, and Rajesh Kaluri. "2p3fl: A novel approach for privacy preserving in financial sectors using flower federated learning." Computer Modeling in Engineering & Sciences 140, no. 2 (2024): 2035.
[12] Wang, Zhongyi, Jin Xiao, Lu Wang, and Jianrong Yao. "A novel federated learning approach with knowledge transfer for credit scoring." Decision Support Systems 177 (2024): 114084.
[13] Gajula, Sreenivasulu. "Federated intelligence in financial ecosystems: A privacy-preserving AI framework for cross-border risk analysis." Journal of Information Systems Engineering and Management 10 (2025): 1040-1048.
[14] Jiang, Xiaoxiao, Wenbo Liu, and Boyang Dong. "FedRisk A Federated Learning Framework for Multi-institutional Financial Risk Assessment on Cloud Platforms." Journal of Advanced Computing Systems 4, no. 11 (2024): 56-72.
[15] Martin-Melero, Inigo, Raul Gomez-Martinez, Maria Luisa Medrano-Garcia, and Felipe Hernandez-Perlines. "Comparison of sectorial and financial data for ESG scoring of mutual funds with machine learning." Financial Innovation 11, no. 1 (2025): 84.

ФИНАНСОВАЯ КИБЕРПРЕСТУПНОСТЬ В РЕСПУБЛИКЕ КАЗАХСТАН: УГРОЗЫ И МЕТОДЫ ЗАЩИТЫ

Степаненко Ж.О., Муканова Ж.А.
e-mail: stepanenkozasmin@gmail.com
Университет «Туран», Республика Казахстан

Аннотация: *В статье рассматриваются ключевые угрозы для финансового сектора и БВУ Казахстана, возникшие в условиях ускоренной цифровизации и внедрения искусственного интеллекта. Проанализированы актуальные формы цифровых преступлений: AI-фишинг, вредоносное ПО, дипфейки, трояны и скрытые дропперские инфраструктуры. Описан комплекс адаптивных инженерно-технических мер защиты: алгоритмы интеллектуального скоринга транзакций (RLS, APC-III, Causal ML), архитектура Zero Trust, стандартизация защиты Open API (mTLS, RASP, MTD), беспарольная аутентификация FIDO2/WebAuthn с контролем «живости» и постквантовая криптография. Предложены образовательные симуляторы атак и институт веб-констеблей для обеспечения стабильности банковского сектора РК.*

Введение

В Республике Казахстан ускоренное развитие цифровой экономики и масштабный переход населения на онлайн-банкинг превратили финансовую киберпреступность, в одну из ключевых угроз экономической стабильности. В 2025-2026 годах особую опасность представляют высокотехнологичные атаки с применением искусственного интеллекта (AI phishing), мультимодальных дипфейков (подделка аудио и видео в реальном времени), троянов удаленного доступа (RAT-платформы).

Теоретическую основу исследования составили труды зарубежных и казахстанских экспертов [1-10]. Оценка академического интереса к теме (по данным Google Scholar по поисковому запросу «Cybersecurity in the financial sector») показывает устойчивый рост публикаций в период с 2021 г. по 2025 г. (рисунок 1).

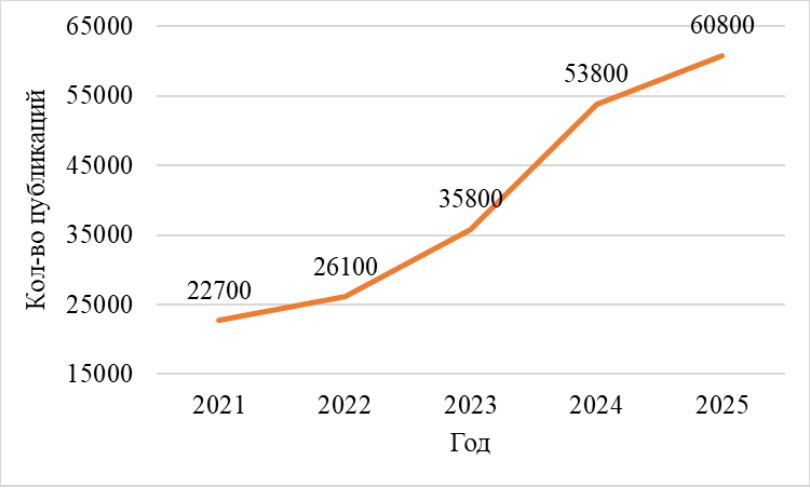


Рисунок 1. Динамика научных публикаций по теме кибербезопасности (Google Scholar, 2021 - 2025 гг.)

Цель исследования – комплексный анализ специфики финансовой киберпреступности в Казахстане, оценка системных уязвимостей банков второго уровня (БВУ) и формирование эшелонированной системы защиты.

Задачи исследования:

1) систематизировать основные виды киберпреступлений в РК;

2) выявить факторы уязвимости на программном, инфраструктурном и человеческом уровнях;

3) предложить адаптивные инженерно-технические и образовательные методы защиты с учетом международного опыта.

1. Анализ киберугроз и уязвимостей финансового сектора РК

Национальная финтех-инфраструктура РК имеет локальную специфику: доминирование универсальных банковских суперприложений (Kaspi Bank, Halyk Bank, Freedom Bank) расширяет область потенциальных атак. Основные типы угроз и их технические параметры систематизированы в Таблице 1.

Таблица 1. Основные виды киберпреступлений и их технические аспекты

Вид киберпреступления	Техническая реализация	Цель атаки	Уровень угрозы в РК
AI-driven Phishing &Дипфейки [2][5]	Использование LLM для персонализированного текста; генерация дипфейков (видео и синтез голоса) при видеозвонках	Обход биометрии FaceID, кража одноразовых SMS-кодов и паролей	Критически высокий
Малварь (Вредоносное ПО) [6]	Мобильные (Android/iOS), overlay-атаки (подмена экранов), утилиты удаленного доступа (RAT), RaaS	Перехват управления устройством, автоматический перехват SMS и транзакций	Высокий
Инфраструктурные атаки на API и серверы [4]	Эксплуатация уязвимостей Open API и Supply Chain ИТ-поставщиков; маскирующие DDoS-атаки	Нарушение работы процессинга, скрытое изменение логов БД	Средний / Высокий
Автоматизированные дропперские сети [2][8]	Координация дропов через Telegram-боты; P2P-автоматизация на криптобиржах; виртуальные SIM-карты	Моментальный транзит, расщепление и выводы KZT в криптовалюту (USDT).	Высокий

1.1. Трехуровневая архитектура системных уязвимостей БВУ РК

Исследования Алимжанова Л.М., Спанова Е.М. и Bohdan Haidabrus подтверждают, что 75% нарушений безопасности связаны с социальной инженерией, а 90% утечек данных вызваны ошибками персонала и пользователей [10]. Системный анализ показывает, что ключевые слабости современных финтех-платформ сосредоточены на трех архитектурных уровнях (рисунок 2).

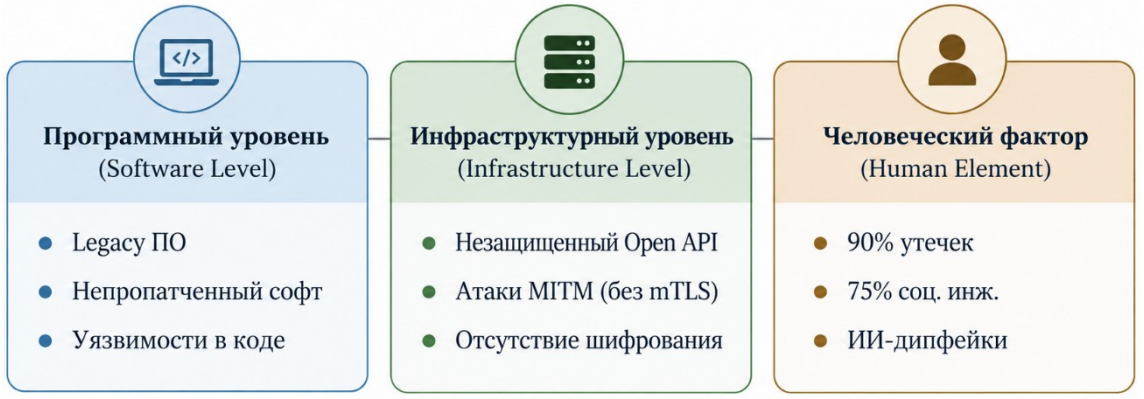


Рисунок 2. Трехуровневая архитектура системных уязвимостей БВУ РК

С точки зрения программной инженерии и системного анализа, бессвязность методов защиты на каждом уровне значительно затрудняет сквозной мониторинг угроз. Ускоренное развитие финтех-экосистем, внедрение новых B2B-серверов и интеграционных API усиливают описанный А. Брассиловой и М. С. Сарыбаевым дефицит потенциала цифровизации (digitalization-capacity) [8]: расширение поверхности атак происходит быстрее, чем развитие средств автоматизированного контроля уязвимостей. Несмотря на действующие нормативные требования и обязательное подключение БВУ к автоматизированной системе обработки информации «Qainar», защита распределенной финтех-инфраструктуры остается уязвимым звеном [3]. Причины и следствия уязвимостей представлены в Таблице 2.

Таблица 2. Системные уязвимости финтех-платформ и их последствия

Архитектурный уровень	Выявленная уязвимость / Слабое место	Техническая первопричина (Root Cause)	Экономические, репутационные и юридические последствия
Программный	Серьезные уязвимости в серверном ПО [4].	Эксплуатация legacy ПО и непропатченных компонентов [4].	Риск RCE-выполнения кода, компрометация баз процессинга.
Инфраструктурный	Незащищенность шлюзов Open API и каналов удаленного доступа [4].	Отсутствие обязательных mTLS и сквозного шифрования трафика с внешними B2B-партнёрами [4].	Перехват данных (MITM), подмена токенов авторизации и мошеннические транзакции.
Человеческий	Высокая восприимчивость пользователей к социальной инженерии [2, 10].	Низкая киберграмотность; обход статической биометрии дипфейками [2, 10].	Добровольная передача реквизитов, вывод средств через дропперские P2P-сети.

2. Программно-технические и инженерные методы защиты

Финансовый сектор Республики Казахстан является одним из самых цифровизованных среди стран СНГ [9]. Активный рост безналичных операций и масштабное развитие банковских суперприложений превращают национальную инфраструктуру в привлекательную цель для высокотехнологичных атак, что требует внедрения методов защиты нового поколения.

2.1. Архитектура нулевого доверия (Zero Trust) и агентская микросегментация в БВУ РК

В центрах обработки данных (ЦОД) казахстанских банков внедряется концепция Zero Trust и Software-Defined Perimeter (SDP) [4]. Специфика инфраструктуры БВУ РК заключается в наличии устаревшего сетевого оборудования (коммуникаторов без поддержки виртуального тегирования). В этих условиях применяется подход агентской изоляции (agent-based isolation): локальные межсетевые экраны устанавливаются непосредственно на хостах, блокируя несанкционированное боковое перемещение (lateral movement) злоумышленников внутри сети [4].

2.2. Интеллектуальный транзакционный скоринг (AI/ML и Causal ML)

Для нейтрализации мошеннических операций в банковском секторе РК стандартные правила заменяются гибридным конвейером ИИ-скоринга (рисунок 3):

- *RLS & APC-III:* Рекурсивный метод наименьших квадратов (коэффициент забывания 0,95 - 0,99) адаптируется к привычкам клиента, а алгоритм APC-III выявляет до 30% скрытых аномалий в офлайн-выборках БВУ [4].
- *Isolation Forest & Dynamic Threshold*: Модели без учителя присваивают балл аномальности [0..1]. При пиковых нагрузках пороги подстраиваются по скользящему окну, а при «холодном старте» новых пользователей применяются статистические пороги популяции [4].
- *Причинный ИИ (Causal ML):* Методы Double Machine Learning и Causal Forests отделяют случайные корреляции от причинно-следственных связей атак [5, 7].

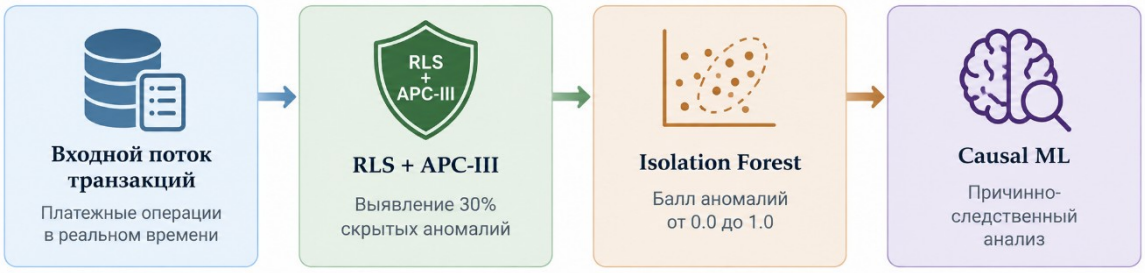


Рисунок 3. Конвейер адаптивного ИИ-скроллинга транзакций БВУ

2.3. Защита Open API, суперприложений и клиентских устройств

Рост B2B-интеграций БВУ с партнерами в рамках Open Banking создает угрозы в цепочке безопасности в цепочке поставок (Supply Chain) [4, 5]. Для их защиты внедряются протоколы взаимной аутентификации mTLS, OAuth 2.0 и механизмы строгой аутентификации (SCA). Мобильные суперприложения защищаются модулями RASP и MTD от банковских троянов (Anatsa и Teabot), overlay-атак и отладки кода [6]. Поведенческая биометрия (Continuous Authentication) обеспечивает непрерывный мониторинг динамики нажатий, скроллинга и наклона смартфона [5][6].

2.4. Беспарольная аутентификация FIDO2 / WebAuthn и постквантовая криптография (PQS).

Для исключения MITM-атак и перехвата SMS OTP банки Казахстана переходят на протоколы FIDO2 и WebAuthn с привязкой к аппаратным модулям (TPM / Secure Enclave) и биометрическим контролем «живости» (liveness detection) [4]. Защита от потенциальных квантовых атак обеспечивается миграцией на решеточную криптографию (CRYSTALS-Kyber, CRYSTALS-Dilithium) с гибридным шифрованием (комбинирование эллиптических и постквантовых ключей через XOR) в TLS-сессиях [4, 6].

3. Образовательные меры и развитие киберграмотности

Поскольку человеческий фактор по-прежнему остается главной причиной большинства инцидентов в сфере информационной безопасности - на него приходится до 90% случаев утечек конфиденциальных данных и свыше 75% успешных атак с использованием социальной инженерии [1, 2, 10]:

- *Интерактивные симуляторы угроз*: Разработка мобильного приложения с симуляцией кибератак (например, фишинговых писем, звонков дипфейк-ботов) поможет обучить пользователей. Это может быть реализовано на Flutter с интеграцией базы данных SQLite для хранения прогресса [1, 2].
- *Национальные просветительские программы:* Интерактивные платформы должны быть интегрированы с государственными проектами, такими как программа «Цифровая грамотность для клиентов банков» целью которой является укрепление уверенности пользователей в цифровых финансовых сервисах и их привыкание к актуальным методам идентификации [1].
- *Внедрение института веб-констеблей*: Специально подготовленные сотрудники правоохранительных органов осуществляют проактивный мониторинг социальных сетей и мессенджеров, своевременно выявляя мошеннические сообщества, фальшивые объявления и предупреждая граждан о новых угроза в реальном времени.

Заключение

Анализ международного опыта показывает, что для снижения рисков несанкционированного межбанковского обмена ключевым фактором является согласование протоколов интеграции с концепцией Open Banking [1]. Отсутствие сквозного шифрования каналов передачи данных и незащищенность публичных API создают риски атак класса MITM и перехвата токенов авторизации [4]. Для казахстанских БВУ стандартизация защиты API-интерфейсов на базе модулей OAuth 2.0 и mTLS поможет ликвидировать уязвимости на границах распределенных экосистем с внешними финтех-партнерами [4]. Хотя реализация этих мер нуждается в значительных инфраструктурных инвестициях, их комплексное внедрение значительно укрепляет долгосрочную стабильность и защищенность транзакционных каналов финансового сектора.

Финансовая киберпреступность в Казахстане представляет собой комплексную, динамически развивающуюся угрозу, связанную с ростом цифровых сервисов, безналичных платежей и интеграционных Open API [1, 8]. Анализ выявил ключевые уязвимости финтех-платформ на трех архитектурных уровнях: программном, инфраструктурном, и человеческом [2, 4, 8].

Предложенные решения, такие как разработка алгоритмов анализа транзакций на базе ИИ и проектирование интерактивных образовательных приложений, могут быть реализованы студентами и молодыми специалистами IT-направлений. Успех зависит от интеграции технических, правовых и образовательных мер, а также от международного опыта. Эти шаги усилят цифровую безопасность Казахстана, обеспечив долгосрочную стабильность финансово-банковского сектора.

ЛИТЕРАТУРА:

1. *G.E. Maulenkulova, A.B. Makhatova, T.N. Mashirova, Y.B. Alimbek*. Digital banking transformation: a bibliometric analysis of global research trends. С. 5-13. - [Электронный ресурс] https://sua.aesa.kz/main/article/view/437/342 (дата обращения 28.08.2026)

2. *ISSAYEVA Zh.A., MUSSABEKOVA I.T., KARZHAUBAYEVA L.M., SANIYAZOVA E.K.*. Phishing in the criminal law of the republic of kazakhstan: legislative gaps and law enforcement challenges. С. 247-253. - [Электронный ресурс] https://esjl.turan-edu.kz/jour/article/view/847/183 (дата обращения 01.09.2026)

3. *T.B. Mussilimova, S.B. Makysh, A.T. Jumabekova*. Banking financial security risks in Kazakhstan. С. 193-202. - [Электронный ресурс] https://bulecon.enu.kz/index.php/main/article/view/1263/750 (дата обращения 01.09.2026)

4. *Yejiao Zhang*. The financial security model of Kazakhstan's second-tier banks in the digital era. С. 1-5 - [Электронный ресурс] https://www.atlantic-press-journals.com/index.php/IJETR/article/view/1140/2032 (дата обращения 03.09.2026)

5. *Sana Ullah Khan, Umair Zafar, Shamikh Imran, Muhammad Ali*. Artificial intelligence–driven fraud detection in fintech: strengthening cybersecurity against digital financial scams. С. 72-79. - [Электронный ресурс] https://insightfuljournals.com/index.php/JBII/article/view/116/191 (дата обращения 03.09.2026)

6. *Małgorzata Such-Pyrgiel, Dariusz Prokopowicz, Anna Mazurkiewicz-Pizło, Anna Gołębiowska, Ewa Listopadzka*. The growing importance of cybersecurity for mobile banking systems and the use of artificial intelligence in identifying cyber threats. С. 208-235. - [Электронный ресурс] https://www.jomswsge.com/pdf-224805-142840?filename=The-growing-importance-of.pdf (дата обращения 04.09.2026)

7. *Md Sazzad Hossain, Md Khalilor Rahman, Sayem Ul Haque, Kazi Abu Jahed, Dil Tabassum Subha, Md Fazlul Huq Mithu, Mahuma Akter, Sumaiyara Islam Oysee, Md Redwan Laskar, Md Mosheur Rahman*. Causal machine learning for financial crime attribution: uncovering hidden cause-and-effect relationships among fraud, money laundering, and cybersecurity incidents in blockchain ecosystems. С. 435-450. - [Электронный ресурс] https://cspub-ijcisim.org/index.php/ijcisim/article/view/2738/2209 (дата обращения 05.09.2026)

8. *Brassilova A., Sarybayev M.S.*. Transnational financial fraud in the context of digitalization: institutional challenges of Kazakhstan and international countermeasures. С. 91-107 - [Электронный ресурс] https://www.researchgate.net/profile/Meiram-Sarybayev/publication/408246867_TRANSNATIONAL_FINANCIAL_FRAUD_IN_THE_CONTEXT_OF_DIGITALIZATION_INSTITUTIONAL_CHALLENGES_OF_KAZAKHSTAN_AND_INTERNATIONAL_COUNTERMEASURES/links/6a4382b32267bf198a6a4196/TRANSNATIONAL-FINANCIAL-FRAUD-IN-THE-CONTEXT-OF-DIGITALIZATION-INSTITUTIONAL-CHALLENGES-OF-KAZAKHSTAN-AND-INTERNATIONAL-COUNTERMEASURES.pdf (дата обращения 05.09.2026)

9. *Л.Т. КОПБАЕВА, Л.С. НУРПЕИСОВА, Г.А. САЙМАГАМБЕТОВА, А.И. ЕСТУРЛИЕВА*. цифровые технологии в финансовой сфере Казахстана. С. 228-237. - [Электронный ресурс] https://vestnik.turan-edu.kz/jour/article/view/3854/1665 (дата обращения 07.09.2026)

10. *L.M. Alimzhanova, E.M. Panova, Bohdan Haidabrus*. Risks and threats in the financial sector of Kazakhstan. С. 61-67 - [Электронный ресурс] https://journal.iitu.edu.kz/index.php/ijict/article/view/319/387 (дата обращения 07.09.2026)

АУДИО ДИПФЕЙКТЕРДІ АНЫҚТАУ ҮШІН ДЕРЕКТЕР ЖИЫНТЫҚТАРЫН ЗЕРТТЕУ ЖӘНЕ НЕЙРОНДЫҚ ЖЕЛІ НЕГІЗІНДЕ ЖІКТЕУ

Ә. Т. Турарбек[1], Ш. Темиргазиева[1]
*e-mail: shopiktem@gmail.com
[1]*Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан*

Аңдатпа. *Бұл жұмыста Mel-спектрограмма мен конволюциялық нейрондық желі негізінде аудио дипфейктерді анықтау әдісі қарастырылды. Зерттеу үшін In The Wild Audio Deepfake деректер жиынындағы bona-fide және spoof аудиожазбалары алынды. Аудио сигналдар 16 кГц жиілікте өңделіп, 4 секундтық сегменттерге келтірілді және 128 × 126 өлшемді Mel-спектрограммаларға түрлендірілді. Жіктеу үшін төрт конволюциялық блоктан және Global Average Pooling қабатынан тұратын CNN моделі құрылды. Модельді бағалау кезінде сөйлеушілер оқыту және тест жиындарында қайталанбайтын speaker-disjoint бөлу қолданылды. Тест нәтижесінде accuracy 91,81%, ал macro F1-score 0,9169 болды. Алынған көрсеткіштер CNN моделінің шынайы және жасанды сөйлеу үлгілерін ажырату мүмкіндігін көрсетті.*

Кілттік сөздер: *аудио дипфейк, конфолюциялық нейрондық желі, Mel-спектрограмма, генеративті сөйлеу.*

Кіріспе

Қазіргі цифрлық ортада киберқауіпсіздік тек ақпараттық жүйелерді техникалық қорғаумен шектелмей, цифрлық контенттің түпнұсқалығын қамтамасыз ету мәселесін де қамтиды. Осы тұрғыдан алғанда, аудио deepfake технологиялары жаңа қауіп факторларының біріне айналып отыр. Жасанды түрде қалыптастырылған немесе түрлендірілген дауыс нақты адамның сөйлеу мәнерін жоғары деңгейде имитациялай алатындықтан, оны дауыс арқылы сәйкестендіру жүйелерін алдау, жалған ақпарат тарату және әлеуметтік инженерияға негізделген шабуылдарда пайдалану қаупі бар. Сондықтан аудионың шынайылығын автоматты түрде тексеру және жасанды сөйлеуді дер кезінде анықтау киберқауіпсіздік міндеттерінің бірі ретінде қарастырылады.

Генеративті жасанды интеллекттің қарқынды дамуы шынайы синтетикалық сөйлеуді жасауға мүмкіндік берді. Қазіргі заманғы мәтінді сөйлеуге айналдыру және дауысты түрлендіру технологиялары адам дауысының маңызды сипаттамаларын, соның ішінде айтылымды, интонацияны және сөйлеу стилін қайталай алады. Генеративті сөйлеудің сапасы жақсарған сайын, түпнұсқа және жасанды түрде жасалған жазбаларды ажырату қиындай түседі [1], [2].

Аудио дипфейктер қауіпсіздік пен ақпаратты тексерудің маңызды мәселесіне айналды, себебі синтетикалық сөйлеуді белгілі бір сөйлеушіні имитациялау немесе шынайы болып көрінетін аудио мазмұнды жасау үшін пайдалануға болады. Бұл мәселе әсіресе сандық байланыста, медианы тексеруде және дауысқа негізделген биометриялық жүйелерде өзекті. Мұндай әдістердің дамуы оқыту және бағалау үшін пайдаланылатын деректерге қатты байланысты. Қолданыстағы деректер жиынтықтары динамиктері, жазу шарттары, тілдері, дыбыс сипаттамалары және жалған үлгілерді жасау үшін қолданылатын әдістер бойынша ерекшеленеді.

Аудио сигналдарды қандай әдіспен анықтау маңызды мәселелердің бірі болып табылады. Дыбысты бастапқы күйіндегі толқын түрінде тікелей өңдеудің орнына, аудиожазбаны уақыт пен жиілік арасындағы байланысты көрсететін арнайы форматқа түрлендіруге болады. Солардың бірі — Mel-спектрограммасы. Ол дыбыстағы спектрлік энергияның уақыт бойынша қалай өзгеретінін көрсетеді және оны нейрондық желілер арқылы талдауға ыңғайлы етеді. Сонымен қатар, бұған дейін жүргізілген зерттеулер спектрограмма негізіндегі белгілерді конволюциялық нейрондық желілермен (CNN)

бірге қолдану синтетикалық немесе қолдан жасалған сөйлеуді анықтауда тиімді нәтиже беретінін көрсеткен [3].

Бұл зерттеуде аудио дипфейктерді анықтау үшін In The Wild Audio Deepfake деректер жиыны пайдаланылды. Шынайы (bona-fide) және жасанды (spoof) сөйлеу жазбалары алдын ала өңделіп, Mel-спектрограммаларға түрлендірілді. Алынған уақыт–жиілік сипаттамалары конволюциялық нейрондық желінің кірісі ретінде қолданылып, модель аудионың шынайы немесе жасанды екенін анықтайтын екілік жіктеуді орындады.

Зерттеудің негізгі мақсаты — Mel-спектрограммаға негізделген CNN моделінің аудио дипфейктерді анықтаудағы тиімділігін бағалау.

Алдыңғы зерттеулерге шолу. Аудио дипфейктерді анықтау бағытындағы зерттеулер сөйлеуді синтездеу және дауыс түрлендіру технологияларының дамуымен қатар қарқынды дамып келеді. Қазіргі жұмыстар бір-бірінен аудионы ұсыну тәсілі, қолданылатын классификация моделі және модельді оқыту мен бағалауға алынған деректер жиынтығы бойынша ерекшеленеді. Соңғы зерттеулер бұл салада дәстүрлі акустикалық белгілерден терең нейрондық желілер мен алдын ала оқытылған сөйлеу модельдеріне қарай бетбұрыс болғанын көрсетеді. Сонымен қатар, модельдің оқыту кезінде кездеспеген жаңа синтетикалық аудионы дұрыс анықтауы әлі де өзекті мәселе болып отыр [4], [5].

Аудио дипфейктерді зерттеуде ASVspoof деректер жиыны кеңінен пайдаланылады. ASVspoof 2019 нұсқасында TTS және дауыс түрлендіру (VC) технологиялары арқылы алынған синтетикалық сөйлеу үлгілері берілген. Оның Logical Access (LA) бөлігі жасанды жолмен қалыптастырылған сөйлеуді зерттеуге бағытталған. Деректерде әртүрлі синтездеу және түрлендіру әдістерінің болуы анықтау модельдерін салыстырмалы түрде бағалауға мүмкіндік береді [6].

Синтетикалық сөйлеуді анықтауға арналған WaveFake деректер жиыны бірнеше нейрондық TTS жүйелері арқылы жасалған аудио үлгілерді қамтиды. Жиынтықта ағылшын және жапон тілдеріндегі жазбалар бар [7]. Әртүрлі генерациялау тәсілдерінің қолданылуы модельдің синтетикалық сөйлеуге тән белгілерді анықтау мүмкіндігін зерттеуге жағдай жасайды. Сонымен бірге мұндай деректерде жоғары нәтиже көрсеткен модель басқа генераторлардан алынған немесе нақты ортадағы аудиоларда төмен нәтиже көрсетуі мүмкін.

Zhou және Lim аудио дипфейктерді тек дыбыстық ақпарат арқылы емес, бейне деректерімен бірге қарастыруды ұсынды [8]. Олардың әдісінде сөйлеу мен адамның бет қимылы арасындағы сәйкестік талданып, осы екі арнада пайда болатын қайшылықтар дипфейкті анықтау белгісі ретінде пайдаланылады. Зерттеу нәтижелері аудио және визуалды ақпаратты біріктіру кейбір жағдайларда тек бір модальділікке сүйенетін модельдерге қарағанда тиімдірек болатынын көрсетті. Әсіресе модельді оқыту кезінде кездеспеген deepfake үлгілерін анықтауда мұндай тәсілдің артықшылығы байқалған. Бұл жұмыс аудио дипфейкті анықтауда дыбыстық белгілерді жеке қарастырумен қатар, оларды басқа модальділіктермен байланыстыра талдаудың да мүмкін екенін көрсетеді [8].

Зерттеулердің тағы бір бағыты аудио сигналды нейрондық желіге спектрограмма түрінде беруге негізделген. Бұл жағдайда бастапқы дыбыс толқыны тікелей өңделмей, алдымен уақыт пен жиілік арасындағы байланысын көрсететін кескінге түрлендіріледі. Мұндай ұсыну форматы конволюциялық модельдерге аудиодағы жергілікті акустикалық үлгілерді анықтауға мүмкіндік береді. Бұл тәсілдің маңыздылығы синтетикалық сөйлеуде адам құлағына немесе бастапқы waveform-ға бірден байқала бермейтін ұсақ спектрлік ерекшеліктердің кездесуімен байланысты. Зерттеулерде синтетикалық және spoofed speech анықтау үшін CNN, residual, recurrent және attention негізіндегі әртүрлі архитектуралар қолданылғаны көрсетіледі [4], [5].

Benchmark датасеттеріндегі жоғары нәтижелерге қарамастан, модельдің жаңа деректерге бейімделуі әлі де шешімін толық тапқан жоқ. Мысалы, оқыту және тестілеу үлгілері бірдей жағдайларда алынған кезде модель жоғары дәлдік көрсетуі мүмкін. Ал синтездеу әдісі, аудио кодегі, жазу ортасы немесе деректер жиынтығы өзгерген жағдайда оның нәтижесі төмендеуі ықтимал. Зерттеулер мұндай жағдайға датасетке тән ерекшеліктердің модельге шамадан тыс әсер етуін, белгілі бір генератор қалдырған белгілерге артық бейімделуді және деректер арасындағы айырмашылықты негізгі себептердің бірі ретінде қарастырады [5]. Сондықтан датасетті тек модельді оқытуға арналған аудио үлгілерінің жиынтығы ретінде емес, соңғы классификация нәтижесіне ықпал ететін маңызды зерттеу компоненті ретінде қарастырған жөн.

Осы мәселелерді ескере отырып, бұл жұмыста аудио сигналды өңдеу, оны Mel-спектрограмма түрінде ұсыну және CNN арқылы жіктеу кезеңдерінің өзара байланысы қарастырылады.

Аудио дипфейктерді анықтауға арналған деректер жиынтықтары. Аудио дипфейкті анықтау моделінің нәтижесі тек оның архитектурасына емес, сонымен қатар модель үйренетін деректердің ерекшеліктеріне де тәуелді. Қазіргі уақытта қолданылатын датасеттерде синтетикалық сөйлеуді алу тәсілдері, сөйлеушілер мен тілдердің құрамы, сондай-ақ аудионың жазылу жағдайлары әртүрлі. Бұдан бөлек, нақты және жасанды үлгілердің арақатынасы да датасетке қарай өзгереді. Мұндай айырмашылықтар модельдің оқу барысында қандай акустикалық белгілерді меңгеретініне және бұрын кездеспеген аудиоларды өңдеу кезінде қандай нәтиже көрсететініне ықпал етеді. Сондықтан датасетті таңдауда оның тек көлемін емес, құрылымын және құрамындағы аудио жазбалардың сипаттарын да ескеру қажет.

ASVspoof аудио дипфейктерді анықтауға арналған зерттеулерде кеңінен қолданылатын деректер жиынтығы болып табылады. ASVspoof 2019 Logical Access (LA) бөлігінде TTS және дауыс түрлендіру (VC) жүйелері арқылы жасалған синтетикалық сөйлеу үлгілері қамтылған. Оқыту жиынында 2 580 шынайы және 22 800 жалған, ал development жиынында 2 548 шынайы және 22 296 жалған сөйлеу үлгісі берілген [9]. Әртүрлі генерациялау тәсілдерінің қамтылуы модельдердің жалған сөйлеуді анықтау қабілетін салыстыруға мүмкіндік береді. Кейінгі ASVspoof 2021 нұсқасында DeepFake (DF) бағыты енгізіліп, шынайы және жасанды сөйлеуді анықтауға арналған деректер ұсынылды [10].

WaveFake датасеті синтетикалық сөйлеуді анықтау міндетіне арнайы бейімделіп жасалған. Оның құрамында алты түрлі нейрондық сөйлеу генерациялау архитектурасы арқылы алынған 117 985 синтетикалық аудио үлгісі бар. Олардың қатарында MelGAN, Parallel WaveGAN, Multi-band MelGAN, Full-band MelGAN, HiFi-GAN және WaveGlow модельдері қолданылады. Датасет LJSpeech және JSUT корпустарына негізделген, сондықтан онда ағылшын және жапон тілдеріндегі сөйлеу жазбалары қамтылған. Бірнеше генерация архитектурасынан алынған үлгілердің болуы анықтау моделін әртүрлі синтездеу тәсілдері жағдайында тексеруге мүмкіндік береді. Бұл WaveFake-ті синтетикалық сөйлеудің әртүрлі генераторларынан пайда болатын акустикалық ерекшеліктерді зерттеу үшін қолайлы деректер жиынтығына айналдырады [11].

Fake-or-Real (FoR) датасеті нақты және синтетикалық сөйлеуді екі класқа бөліп, оларды ажырату міндетіне бағытталған. Оның құрамында шамамен 111 000 нақты және 87 000 жалған аудио жазба бар. Аудиолар әртүрлі ұзақтықта беріліп, WAV және MP3 форматтарын қамтиды. Жалған сөйлеу үлгілерін алу үшін Deep Voice 3 және Google WaveNet сияқты мәтінді сөйлеуге түрлендіру жүйелері қолданылған. FoR-дың басты ерекшелігі – классификация міндетінің нақты екі класқа құрылуы: модель аудионың шынайы немесе жасанды екенін анықтайды. Осы құрылымы оны CNN сияқты нейрондық желілерді қолданып, нақты және синтетикалық сөйлеуді ажырату эксперименттеріне ыңғайлы етеді [12].

FakeAVCeleb датасеті дипфейктерді тек аудио арқылы емес, аудио және бейне ақпараттарын қатар пайдаланып зерттеуге мүмкіндік береді. Датасетте барлығы 20 000 бейнежазба қамтылған, оның 500-і нақты, 19 500-і жалған бейнелер. Сонымен бірге бейнежазбаларға сәйкес нақты және синтетикалық аудио үлгілері де ұсынылған. Мұндай құрылым сөйлеудегі дыбыстық ақпарат пен бейнедегі адамның қимыл-қозғалысы арасында пайда болатын сәйкессіздіктерді зерттеуге мүмкіндік береді [13].

ADD 2022 бұрынғы сөйлеуді бұрмалауды анықтауға арналған деректер жиындарында толық қамтылмаған күрделі жағдайларды зерттеу үшін жасалған. Бұл деректер жиыны үш бағытты қамтиды: сапасы төмен жалған аудионы анықтау, аудионың белгілі бір бөлігінің ғана өзгертілгенін анықтау және жалған аудионы ажыратуға арналған арнайы тапсырма [14].

In-the-Wild датасеті модельдің тек арнайы дайындалған деректерде емес, күнделікті қолданыста кездесетін аудиоларды да қаншалықты дұрыс ажыратынын тексеру үшін ұсынылған. Онда интернеттегі ашық дереккөздерден жиналған 58 танымал адамның нақты және жалған сөйлеу жазбалары бар. Датасетте шамамен 20,8 сағат нақты (bona fide) және 17,2 сағат жалған (spoofed) аудио қамтылған [15].

Соңғы жылдары аудио дипфейктерді анықтауда көптілді деректерді қолдануға да көңіл бөлінуде. Осы бағыттағы MLAAD (Multilingual Audio Anti-Spoofing Dataset) датасеті 23 тілдегі синтетикалық сөйлеу үлгілерін қамтиды. Оның жасалуына бұрынғы датасеттерде тілдік әртүрліліктің шектеулі болуы себеп болды. Зерттеулерде MLAAD құрамында шамамен 163,9 сағат синтетикалық сөйлеу бар екені көрсетілген [16].

1-кесте. Аудио дипфейктерді анықтауға арналған деректер жиынтықтарын салыстыру

Деректер жиынтығы	Файлдар/ үлгілер саны	Кластары	Тілдері	Генерация/ өзгерту әдісі	Негізгі мақсаты
ASVspoof 2019	121 461 сөйлеу үлгісі	Нақты / Жалған	Ағылшын	TTS, VC, қайта ойнату	Спуфингті анықтау
ASVspoof 2021 DF	611 829 сөйлеу үлгісі	Нақты / Жалған	Ағылшын	TTS, VC	Дипфейк сөйлеуді анықтау
WaveFake	134 268 аудио файл	Нақты / Жалған	Ағылшын, жапон	Нейрондық TTS, вокодерлер	Синтетикалық сөйлеуді анықтау
Fake-or-Real (FoR)	≈198 000 аудио файл	Нақты / Жалған	Ағылшын	TTS, синтетикалық сөйлеу	Екі класты классификация
FakeAVCeleb	20 000 бейнежазба	Нақты / Жалған	Ағылшын	Аудио-бейне манипуляциясы	Мультимодальды дипфейктерді анықтау
ADD 2022	154 949 аудио үлгі*	Нақты / Жалған / Ішінара жалған	Қытай	TTS, VC, ішінара өзгерту	Аудио дипфейктерді анықтау
In-the-Wild	31 779 аудио файл	Нақты / Жалған	Ағылшын	Нақты ортадағы синтетикалық сөйлеу	Жалпылау қабілетін бағалау
MLAAD	163,9 сағат синтетикалық сөйлеу	Синтетикалық	23 тіл	TTS	Көптілді анти-спуфинг

* ADD 2022 бірнеше тапсырмадан (track) тұрады, сондықтан көрсетілген жалпы сан осы тапсырмалардың үлгілерін біріктіру арқылы алынған.

Салыстыру нәтижесі аудио дипфейктерді анықтаудың барлық аспектілерін толық қамтитын бір ғана деректер жиынтығы жоқ екенін көрсетеді. ASVspoof стандартталған бағалау жүргізуге және әртүрлі спуфинг шабуылдарын зерттеуге мүмкіндік береді, WaveFake бірнеше нейрондық сөйлеу генерациялау архитектураларына негізделген, ал ADD 2022 шулы және ішінара өзгертілген сөйлеу үлгілерін қамтиды. In-the-Wild деректер жиынтығы модельдің нақты ортадағы аудиоларға жалпылау қабілетін бағалауға бағытталған. FakeAVCeleb мультимодальды зерттеулер үшін қолайлы болса, MLAAD алдыңғы деректер жиынтықтарындағы тілдік шектеулерді азайтуға бағытталған.

Зерттеу деректерінің сипаттамасы. Зерттеуде аудионың түпнұсқалығын анықтау үшін In The Wild Audio Deepfake деректер жинағы пайдаланылды. Бұл деректер жиынтығы шынайы және жасанды түрде түрлендірілген сөйлеу үлгілерін қамтиды және аудио deepfake анықтау міндетіне бейімделген. Деректер жинағында әрбір аудио жазбаға оның класын және сөйлеушісін сипаттайтын метадеректер берілген.

Зерттеу барысында бастапқы метадеректер файлында көрсетілген 31 779 аудио жазбаның барлығы өңдеуге енгізілді. Олардың ішінде 19 963 жазба bona-fide, яғни шынайы сөйлеу үлгілеріне, ал 11 816 жазба spoof, яғни жасанды немесе қолдан жасалған сөйлеу үлгілеріне жатқызылды. Деректерде барлығы 54 түрлі сөйлеуші қамтылды. Класстарды белгілеу кезінде bona-fide класына 0, ал spoof класына 1 сандық мәні берілді.

Зерттеудің маңызды ерекшелігі – деректерді оқыту және тестілеу кезінде сөйлеушілердің бір-бірімен араласып кетуіне жол бермеу. Себебі бір адамның дауыс ерекшеліктері оқыту және тест жиындарында қатар кездессе, модель deepfake белгілерінен бөлек, нақты сөйлеушінің акустикалық сипаттарын есте сақтауы мүмкін. Осы ықтимал әсерді азайту мақсатында деректерді бөлу сөйлеуші деңгейінде жүргізілді. Нәтижесінде модельді бағалау кезінде тест жиынында оқыту кезеңінде кездеспеген сөйлеушілердің аудиолары пайдаланылды.

Әрбір аудио жазбаның файл жолы, класы және сөйлеуші идентификаторы жеке сақталып, кейінгі өңдеу кезеңдеріне осы үш ақпарат бірге берілді. Мұндай ұйымдастыру модельдің тек аудио класына ғана емес, деректердің құрылымына да тәуелділігін бақылауға мүмкіндік береді. Деректерді дайындаудың келесі кезеңінде барлық аудио жазбалары бірыңғай акустикалық форматқа келтіріліп, CNN моделіне беруге қолайлы уақыт–жиілік сипаттамасына түрлендірілді.

Аудио сигналдарды алдын ала өңдеу. Аудио жазбалардың ұзақтығы мен бастапқы техникалық параметрлері бірдей болмағандықтан, оларды модельге тікелей беру мүмкін емес. Сондықтан барлық жазбалар бірыңғай форматқа келтірілді. Алдын ала өңдеудің негізгі мақсаты — аудио үлгілерінің өлшемін стандарттау және кейінгі уақыт–жиілік түрлендіруі кезінде олардың өзара салыстырмалылығын сақтау.

Алдымен аудио сигналдардың дискреттеу жиілігі 16 кГц мәніне келтірілді. Егер бастапқы жазбаның дискреттеу жиілігі 16 кГц-тен өзгеше болса, сигнал қайта дискреттелді. Бұл қадам барлық үлгілердегі уақыт бойынша және жиілік бойынша ақпараттың бірдей масштабта берілуін қамтамасыз етеді. Көпарналы аудио жазбалар бір арналы сигналға түрлендірілді. Ол үшін бірнеше арнадағы амплитудалар бойынша орташа мән есептеліп, нәтижесінде моно сигнал алынды. Осылайша модельге бір аудио жазбадан біртекті акустикалық ақпарат беріледі және арналар санына байланысты өлшем айырмашылығы жойылады.

Әрбір аудио жазбаның ұзақтығы 4 секунд ретінде бекітілді. Ұзақтығы 4 секундтан асатын жазбалардан орталық бөліктегі 64 000 үлгі алынды. Орталық сегментті таңдау арқылы сигналдың басындағы немесе соңындағы кездейсоқ бөліктің ғана сақталуына тәуелділікті азайту көзделді. Ал ұзақтығы 4 секундтан қысқа жазбалар үшін сигналдың

соңына нөлдік мәндер қосылып, оның ұзындығы 64 000 үлгіге жеткізілді. Нәтижесінде барлық аудио жазбалар бірдей уақыттық өлшемге ие болды.

Осы өңдеу қадамдарынан кейін әрбір үлгі 16 кГц дискреттеу жиілігімен, моно арналы және 4 секундтық стандартталған аудио сигнал түрінде алынды. Дайындалған сигналдар келесі кезеңде Mel-спектрограммаға түрлендіріліп, CNN моделінің кіріс деректері ретінде пайдаланылды.

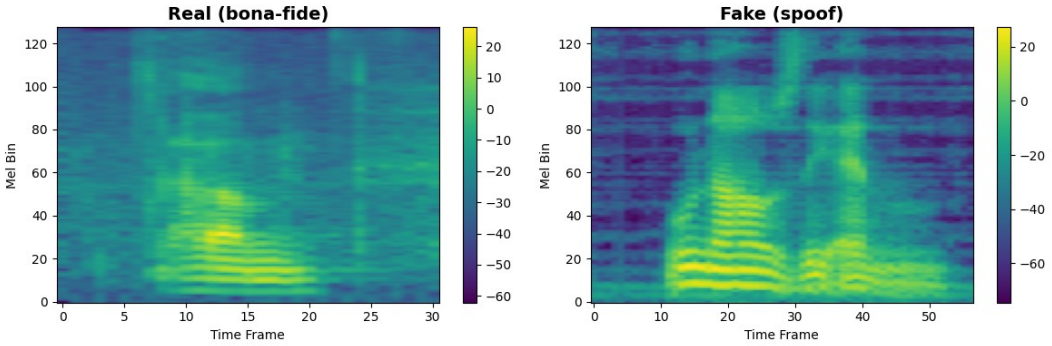
Алдын ала өңделген аудио сигналдар CNN моделіне берілмей, уақыт–жиілік сипаттамаларын алу үшін Mel-спектрограммаға түрлендірілді. FFT өлшемі 1024, hop length 512, Mel-фильтрлер саны 128 болып белгіленді. Нәтижесінде әрбір 4 секундтық аудио үшін өлшемі 128 × 126 Mel-спектрограмма алынды.

Спектрлік мәндер AmplitudeToDB(stype='power') функциясы арқылы децибелдік шкалаға түрлендіріліп, әрбір спектрограмма жеке нормализацияланды:

$$X_{norm} = \frac{X-\mu}{\sigma+\varepsilon},\ \varepsilon = 10^{-6} \quad (1)$$

Мұндағы X - бастапқы Mel-спектрограмма, μ - оның орташа мәні, σ - стандартты ауытқуы, X_{norm} - нормализацияланған спектрограмма, ал $\varepsilon = 10^{-6}$ - нөлге бөлуді болдырмау үшін енгізілген кіші тұрақты шама.

Bona-fide және spoof аудио үлгілерінің Mel-спектрограммалары 1-суретте көрсетілген.



Сурет 1. Bona-fide және spoof аудио сигналдарының Mel-спектрограммалары

Осы өңдеу нәтижесінде 31 779 аудио жазбаның барлығына Mel-спектрограммалар қалыптастырылды. CNN моделінің кіріс өлшемі 1 × 128 × 126 болды.

CNN моделінің архитектурасы. Аудио жазбаларды екіөлшемді Mel-спектрограмма түрінде ұсынғаннан кейін оларды автоматты түрде жіктеу үшін конволюциялық нейрондық желі (CNN) қолданылды. Модельдің кірісіне өлшемі 1 × 128 × 126 болатын бір арналы Mel-спектрограмма берілді. Желінің архитектурасы спектрограммадағы жергілікті уақыт–жиілік ерекшеліктерін біртіндеп бөліп алып, соңында аудио үлгісін екі класстың біріне жатқызуға бағытталған. CNN архитектурасы төрт конволюциялық блоктан тұрады. Әрбір блокта конволюциялық қабат, Batch Normalization, ReLU активация функциясы және 2*2 өлшемді Max Pooling операциясы орналасқан. Конволюциялық қабаттарда белгілер карталарының саны әр блок сайын арттырылып, ал кеңістіктік өлшемдері pooling арқылы біртіндеп азайтылады.

Бірінші блокта кіріс бір арналы спектрограмма 32 белгілер картасына түрлендіріледі. Max Pooling операциясынан кейін деректердің өлшемі 32 × 64 × 63 болады. Екінші блокта белгілер карталарының саны 64-ке дейін көбейіп, өлшем 64 × 32 × 31 түріне өзгереді. Үшінші блокта 128 белгілер картасы қалыптасып, оның өлшемі 128 × 16 × 15 болады. Төртінші блоктан кейін 256 белгілер картасы алынып, кеңістіктік өлшем 256 × 8 × 7 түріне дейін қысқарады. Осылайша желінің қабаттары арқылы өткен сайын

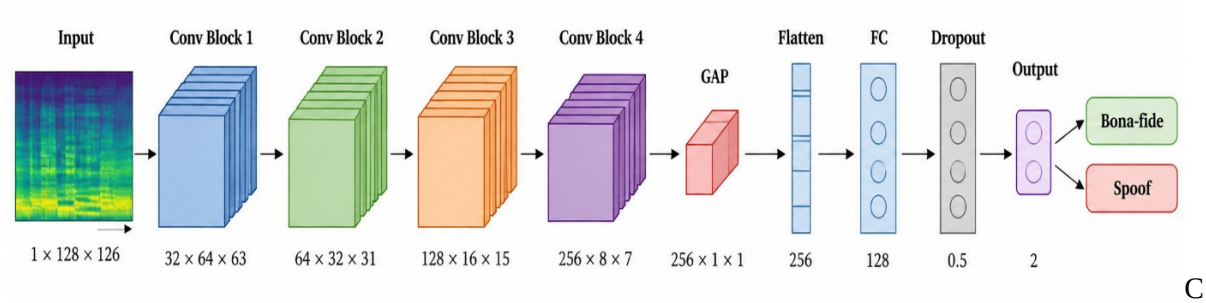
Mel-спектрограммадағы бастапқы уақыт–жиілік ақпарат ықшам белгілер кеңістігіне түрленеді.

Арналар санының 32-ден 256-ға дейін өсуі күрделірек акустикалық сипаттамаларды бөліп алуға мүмкіндік береді, ал Max Pooling операциясы есептеу көлемін азайтып, маңызды белгілерді сақтай отырып кеңістіктік өлшемді қысқартады.

Төртінші конволюциялық блоктан кейін Global Average Pooling қолданылды. Бұл операция әрбір 256 белгілер картасы бойынша кеңістіктік өлшемді бір ғана мәнге дейін жинақтап, нәтижесінде 256 элементтен тұратын вектор қалыптастырады. Одан кейін бұл вектор толық байланысқан қабатқа беріледі.

Толық байланысқан қабатта 256 кіріс нейроны және 64 шығыс нейроны бар. Қабаттан кейін ReLU активациясы және 0,2 ықтималдықпен Dropout қолданылды. Соңғы толық байланысқан қабат 64 өлшемді векторды екі шығыс нейронына түрлендіреді. Бұл екі нейрон сәйкесінше bona-fide және spoof кластарына сәйкес келеді.

Модельдің жалпы оқытылатын параметрлер саны 405 378 болды. Ұсынылған CNN моделінің архитектурасы 2-суретте көрсетілген.



С

урет 2. Аудио дипфейктерді анықтауға арналған CNN моделінің архитектурасы

Осы архитектурада конволюциялық қабаттар Mel-спектрограммадағы акустикалық ерекшеліктерді автоматты түрде бөліп алса, соңғы толық байланысқан қабат алынған белгілер негізінде аудионың bona-fide немесе spoof класына жатуын анықтайды.

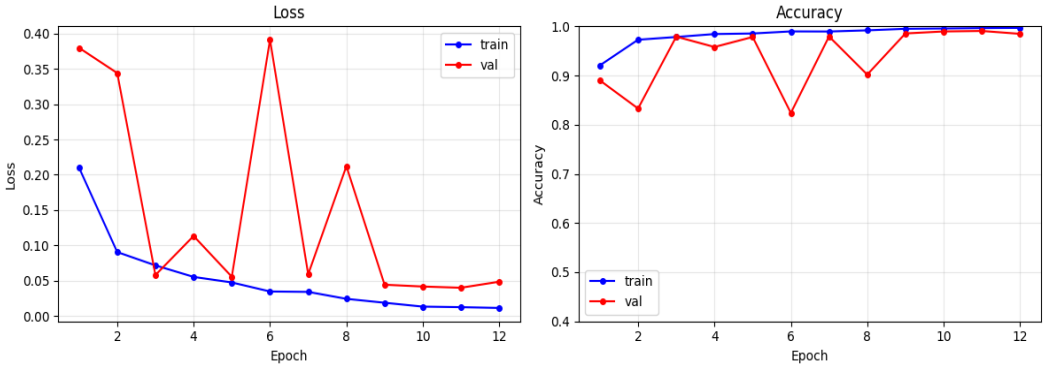
Модельді оқыту стратегиясы. CNN моделін оқыту кезінде класстар арасындағы үлгілер санының айырмашылығын ескеру үшін салмақталған Cross-Entropy Loss қолданылды. Модель параметрлерін жаңарту AdamW оптимизаторы арқылы орындалып, бастапқы learning rate және weight decay мәндері тәжірибеде белгіленген параметрлерге сәйкес алынды. Модель 12 дәуір, batch size 32 режимінде оқытылды. Оқыту процесінің тұрақтылығы үшін CosineAnnealingLR жоспарлаушысы пайдаланылып, кездейсоқтық тұқымы 42 мәніне бекітілді. Есептеулер CUDA қолдайтын GPU арқылы жүргізілді.

Модельді оқыту нәтижелері. CNN моделінің training accuracy көрсеткіші 99,68%-ға жетті. Бұл модельдің Mel-спектрограммалардан шынайы және жасанды аудиоға тән белгілерді тиімді үйренгенін көрсетеді. Модельдің жалпылау қабілетін бағалау үшін оқытуда кездеспеген сөйлеушілерден тұратын speaker-disjoint тест жиыны қолданылды. Тест нәтижесінде accuracy 91,81% болды. Training және тест көрсеткіштерінің айырмашылығы модельдің жаңа сөйлеушілер аудиосын жіктеу кезінде қиынырақ жұмыс істейтінін көрсетеді. Сондықтан модельдің жалпылау қабілетін бағалауда 91,81% тест accuracy негізгі нәтиже ретінде қарастырылады.

Жалпы алғанда, алынған нәтижелер Mel-спектрограммаға негізделген CNN архитектурасының аудио deepfake анықтау міндетінде жоғары жіктеу қабілетіне ие екенін көрсетті. Алайда оқыту және тәуелсіз тест нәтижелері арасындағы айырмашылық сөйлеушілер арасындағы акустикалық ерекшеліктердің модельдің шешіміне әсер етуі мүмкін екенін де көрсетеді.

3-суретте CNN моделін оқыту барысында оқыту (train) және валидациялық (validation) жиындарындағы loss пен accuracy көрсеткіштерінің epoch бойынша өзгеруі берілген. Оқыту процесі барысында train loss біртіндеп төмендеп, модельдің оқыту

деректеріндегі қателігінің азайғанын көрсетеді. Сонымен қатар train accuracy алғашқы epoch-тағы шамамен 92% мәнінен соңғы кезеңде 99%-дан жоғары деңгейге дейін өскен.



Сурет 3. CNN моделін оқыту барысындағы жоғалту және дәлдік көрсеткіштерінің өзгеруі

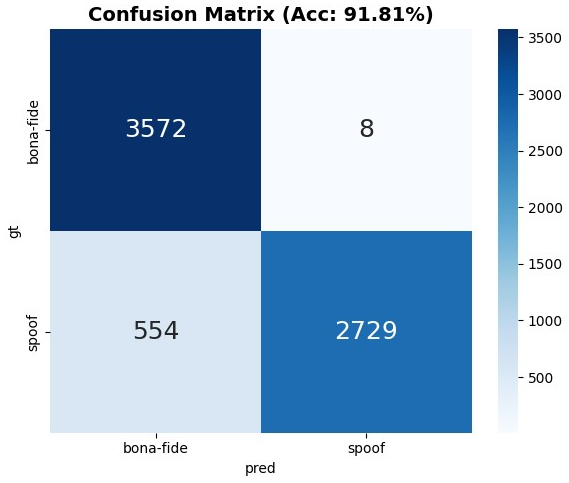
F1-score мәндері bona-fide класы үшін 0,9271, ал spoof класы үшін 0,9066 болды. Екі класс бойынша есептелген macro average нәтижесінде precision 0,9314, recall 0,9145, ал F1-score 0,9169 мәндерін көрсетті.

2-кесте. Тестілік деректер жиынындағы модельдің жіктеу нәтижелері

Class	Precision	Recall	F1-score	Support
bona-fide	0.8657	0.9978	0.9271	3580
spoof	0.9971	0.8313	0.9066	3283
accuracy	—	—	0.9181	6863
macro avg	0.9314	0.9145	0.9169	6863
weighted avg	0.9286	0.9181	0.9173	6863

Нәтижелерді салыстыру модельдің spoof класына қатысты precision көрсеткішінің өте жоғары екенін, ал bona-fide класы бойынша recall көрсеткішінің жоғары екенін көрсетті. Бұл модельдің жалған аудио деп белгіленген үлгілерді негізінен дұрыс анықтайтынын, сонымен қатар шынайы аудиоларды өткізіп алмауға қабілетті екенін көрсетеді. Дегенмен spoof recall көрсеткішінің 0,8313 деңгейінде болуы кейбір жасанды аудио үлгілерінің шынайы ретінде қабылданатынын көрсетеді.

Модельдің жіктеу қателерін егжей-тегжейлі қарастыру үшін confusion matrix қолданылды. Бұл матрица нақты класс пен модель болжаған класс арасындағы сәйкестікті көрсетеді және дұрыс әрі қате жіктелген үлгілердің таралуын бағалауға мүмкіндік береді. Зерттеу жағдайында екі класс қарастырылды: bona-fide және spoof.



Сурет 4. Тестілік жиындағы CNN моделінің қателер матрицасы (Confusion Matrix)

4-суретте CNN моделінің тестілік деректер жиынындағы жіктеу нәтижелері көрсетілген. Жалпы дұрыс жіктелген үлгілер саны 6301 болып, модельдің жалпы дәлдігі 91,81% құрады. Нәтижелер модельдің шынайы сөйлеу үлгілерін анықтауда жоғары көрсеткішке ие екенін, ал жасанды аудио үлгілерінің бір бөлігін шынайы сөйлеу ретінде қате жіктейтінін көрсетеді.

Қорытынды

Бұл зерттеуде аудио deepfake үлгілерін автоматты түрде анықтауға арналған Mel-спектрограмма мен конволюциялық нейрондық желіге негізделген әдіс қарастырылды. Зерттеу барысында In The Wild Audio Deepfake деректер жинағындағы 31 779 аудио жазба өңделіп, олардың 19 963-і bona-fide, ал 11 816-ы spoof класына жатқызылды.

Аудио сигналдар бірыңғай форматқа келтіріліп, 16 кГц дискреттеу жиілігіне түрлендірілді, моно форматқа келтіріліп, ұзақтығы 4 секундқа стандартталды. Әрбір аудио үлгісінен 128 × 126 өлшемді Mel-спектрограмма қалыптастырылып, одан әрі CNN моделінің кіріс деректері ретінде пайдаланылды. Ұсынылған модель төрт конволюциялық блоктан, Global Average Pooling қабатынан және толық байланысқан қабаттардан тұрып, жалпы 405 378 оқытылатын параметрді қамтыды.

Эксперимент барысында модельдің оқыту жиынындағы accuracy көрсеткіші 99,68%, ал валидация жиынындағы ең жоғары accuracy 99,06% болды. Сөйлеушілер бойынша бөлінген тәуелсіз тестілеу кезінде 91,81% accuracy нәтижесі алынды. Осыған байланысты келешектегі зерттеулерде модельдің жаңа сөйлеушілерге жалпылау қабілетін арттыруға, spoof үлгілерін анықтау сезімталдығын жақсартуға және әртүрлі акустикалық жағдайларды қамтитын деректермен эксперименттер жүргізуге назар аудару қажет.

Әдебиеттер

1. X. Wang et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Computer Speech & Language, vol. 64, 2020, Art. no. 101114.

2. J. Frank and L. Schönherr, “WaveFake: A Data Set to Facilitate Audio Deepfake Detection,” arXiv preprint arXiv:2111.02813, 2021.

3. H. Tak et al., “Automatic speaker verification spoofing and deepfake detection using neural networks,” Proceedings of Interspeech, 2021.

4. J. Yi et al., “Audio Deepfake Detection: A Survey,” arXiv preprint arXiv:2308.14970, 2023.

5. M. Li, Y. Ahmadiadli, and X.-P. Zhang, “A Survey on Speech Deepfake Detection,” ACM Computing Surveys, vol. 57, no. 7, 2025.

6. X. Wang et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Computer Speech & Language, vol. 64, 2020, Art. no. 101114.

7. J. Frank and L. Schönherr, “WaveFake: A Data Set to Facilitate Audio Deepfake Detection,” arXiv preprint arXiv:2111.02813, 2021.

8. Y. Zhou and S.-N. Lim, “Joint Audio-Visual Deepfake Detection,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14800–14809, 2021.

9. X. Wang et al., “ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech,” Computer Speech & Language, vol. 64, 2020.

10. J. Yamagishi et al., “ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection,” Proceedings of the ASVspoof 2021 Workshop, 2021.

11. J. Frank and L. Schönherr, “WaveFake: A Data Set to Facilitate Audio Deepfake Detection,” arXiv preprint arXiv:2111.02813, 2021.

12. M. Abdeldayem, “Audio Deep Fake Detection with Sonic Sleuth Model,” Computers, vol. 13, no. 10, 2024.

13. H. Khalid, S. Tariq, M. Kim, and S. S. Woo, “FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset,” NeurIPS Datasets and Benchmarks Track, 2021.

14. J. Yi et al., “ADD 2022: The First Audio Deep Synthesis Detection Challenge,” Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9216–9220, 2022.

15. N. M. Müller et al., “Does Audio Deepfake Detection Generalize?” Proceedings of Interspeech, 2022.

16. J.-W. Jung et al., “MLAAD: The Multi-Language Audio Anti-Spoofing Dataset,” 2024.

ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ СВОЙСТВ КОМПОЗИТНЫХ МАТЕРИАЛОВ НА ОСНОВЕ МНОГОКРИТЕРИАЛЬНОГО ПРИНЯТИЯ РЕШЕНИЙ

Ташатов Н.Н.[1], Жүзбаев С.С.[2], Сарсенбай М.С.[2]
e-mail: tash.nur@mail.ru, zhuzbayev_ss@enu.kz
[1]НИИ информационной безопасности и криптологии, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан
[2]Кафедра информационных систем, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан

Аннотация

Рассматривается обеспечение информационной безопасности информационно-аналитической системы (ИАС) прогнозирования свойств композитных материалов, обучаемой на данных испытаний с IIoT-оборудования. Предложена архитектура доверенного контура данных на основе распределённого реестра, а выбор протокола консенсуса формализован как многокритериальная задача (метод анализа иерархий, TOPSIS, VIKOR). Для консорциума лабораторий предпочтительны PBFT и Tendermint; вывод устойчив к возмущению оценок. Эксперимент показал, что целевая фальсификация 20 % записей почти не меняет R^2, но завышает прогноз прочности атакуемого материала на 12,7 %; криптографический контроль целостности его устраняет.

Ключевые слова: *информационная безопасность, информационно-аналитическая система, композитные материалы, многокритериальное принятие решений, протокол консенсуса, блокчейн, целостность данных, отравление данных.*

Введение

Прогнозирование свойств композитных материалов методами машинного обучения сокращает объём дорогостоящих натурных испытаний при подборе состава, схемы армирования и технологии изготовления [1]. Информационно-аналитическая система (ИАС), реализующая такое прогнозирование, агрегирует результаты испытаний от испытательных машин и датчиков технологического процесса (промышленный Интернет вещей, IIoT), партнёрских лабораторий и литературных источников. Качество прогноза при этом определяется не только алгоритмом, но и достоверностью данных, которая относится к ключевым проблемам материаловедения, основанного на данных [2].

Наиболее опасной угрозой для ИАС этого класса является скрытое искажение данных: преднамеренное изменение результатов испытаний («отравление» обучающей выборки) смещает прогноз модели и может привести к выбору материала с завышенной расчётной прочностью. Систематика таких атак приведена в отчёте NIST [3], уязвимость регрессионных моделей к отравлению показана в [4]. Распределённые реестры позволяют фиксировать неизменяемые отпечатки данных без доверенного посредника и применяются для защиты IoT-систем [5], однако их свойства определяются протоколом консенсуса, выбор которого при противоречивых требованиях к безопасности, производительности и ресурсоёмкости является многокритериальной задачей [6–9].

Цель работы — разработать подход к обеспечению информационной безопасности ИАС прогнозирования свойств композитных материалов, в котором механизм обеспечения целостности данных (протокол консенсуса реестра) выбирается методами многокритериального принятия решений (МКПР), и оценить влияние нарушения целостности данных на качество прогноза.

1. Модель угроз и требования безопасности

Анализ компонентов ИАС по модели STRIDE [10] позволил выделить угрозы, влияющие на достоверность прогноза (таблица 1). Определяющими являются подмена данных испытаний и отрицание авторства записей, поскольку метрики качества модели их не выявляют. Отсюда следуют требования к доверенному контуру данных: криптографическая привязка каждой записи к источнику, неизменяемость зафиксированных отпечатков при взаимном недоверии участников консорциума, конфиденциальность рецептур и технологических режимов, учёт ограниченных ресурсов шлюзов IIoT.

Таблица 1

Угрозы компонентам ИАС и меры защиты

Компонент ИАС	Угроза (STRIDE) и последствие	Мера защиты
Испытательное оборудование, датчики (IIoT)	Подмена источника (S): ложные измерения от имени устройства	Аутентификация устройств, подпись записей ECDSA
Хранилище данных испытаний	Модификация записей (T): скрытое смещение обучающей выборки и прогноза	SHA-256, дерево Меркла, фиксация корня в реестре
Журнал загрузки данных	Отрицание авторства (R): источник искажения не устанавливается	Подписанные транзакции, неизменяемый журнал реестра
Рецептуры и технологические режимы	Раскрытие (I): утечка ноу-хау участников консорциума	Шифрование AES-GCM, хранение вне реестра
Узлы реестра, модуль прогнозирования	Отказ в обслуживании (D), повышение привилегий (E)	BFT-консенсус, разграничение доступа RBAC, аудит

2. Архитектура доверенного контура данных

Предлагаемая архитектура (рисунок 1) разделяет ИАС на три контура. Шлюз доверия аутентифицирует устройство, приводит запись испытания к каноническому виду, вычисляет хеш SHA-256, подписывает запись по ECDSA и объединяет хеши пакета из m записей в дерево Меркла [11]. Записи хранятся в зашифрованном виде вне реестра, а в реестр консорциума заносятся только корни деревьев, метаданные и журнал доступа, поэтому на пакет приходится одна транзакция и требования к пропускной способности реестра невелики.

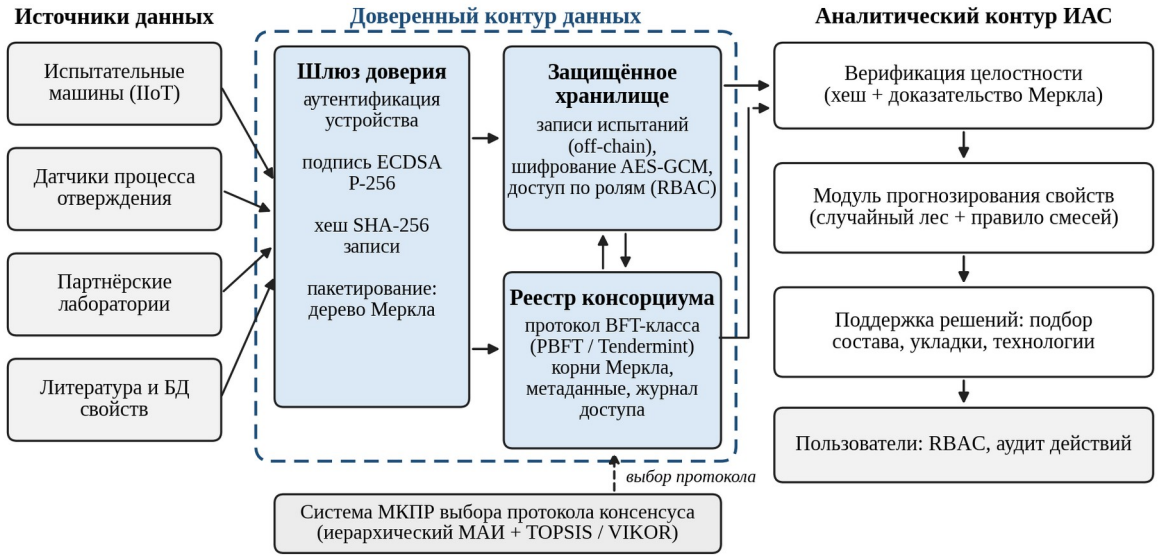


Рисунок 1 – Архитектура ИАС прогнозирования свойств композитных материалов с доверенным контуром данных

Перед обучением модели и при обработке запросов аналитический контур проверяет каждую запись d_k по правилу

$$V(d_k) = [\mathrm{Root}(H(d_k), \pi_k) = R_b], \tag{1}$$

где H — хеш-функция SHA-256; π_k — доказательство включения длиной $\lceil \log_2 m \rceil$ хешей; Root — восстановление корня дерева; R_b — корень пакета b, зафиксированный в реестре; $[\cdot] = 1$ при выполнении условия и 0 в противном случае. Записи с $V(d_k) = 0$ исключаются из обучающей выборки. Модуль прогнозирования основан на случайном лесе [12] с физически обоснованным признаком — оценкой прочности по правилу смесей.

3. Многокритериальная модель выбора протокола консенсуса

Множества альтернатив и критериев приняты в соответствии с методикой и программной системой, разработанными в рамках проекта AP23487259 [6–9]: A_1 — PoS, A_2 — DPoS, A_3 — Raft [13], A_4 — PBFT [14], A_5 — PoA [15], A_6 — Tendermint [16]. Критерии образуют четыре группы: производительность (C_1 — пропускная способность, C_2 — задержка транзакций, C_3 — финализация), ресурсоэффективность (C_4 — энергопотребление, C_5 — вычислительные ресурсы, C_6 — хранилище), безопасность (C_7 — отказоустойчивость, C_8 — стойкость к атакам), применимость (C_9 — децентрализация, C_{10} — управляемость, C_{11} — эксплуатационные затраты). Оценки x_{ij} выставлены экспертно по шкале 1–10 (10 — лучшее значение) на основе [9, 13–16], поэтому все критерии максимизируются.

Веса критериев определяются иерархическим методом анализа иерархий (МАИ) [17]: для матриц парных сравнений групп и критериев внутри групп находятся главные собственные векторы, согласованность проверяется по отношению CR, а итоговый вес равен произведению весов:

$$CR = \frac{\lambda_{\max} - n}{(n-1)RI} \le 0{,}1, \qquad w_j = W_g \cdot w_j^{(g)}, \tag{2}$$

где RI — случайный индекс; W_g — вес группы g; $w_j^{(g)}$ — локальный вес критерия j. Для сценария ИАС, в котором участники консорциума не доверяют друг другу, а интенсивность поступления данных невелика, наибольший приоритет присвоен группе «Безопасность».

Ранжирование выполняется методом TOPSIS [18]: после векторной нормализации $r_{ij} = x_{ij}/(\Sigma_i x_{ij}^2)^{1/2}$ и взвешивания $v_{ij} = w_j r_{ij}$ определяются идеальное $v_j^* = \max_i v_{ij}$ и антиидеальное $v_j^- = \min_i v_{ij}$ решения, расстояния до них и коэффициент близости

$$D_i^* = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^*\right)^2}, \qquad C_i = \frac{D_i^-}{D_i^* + D_i^-}, \tag{3}$$

где D_i^- вычисляется аналогично с заменой v_j^* на v_j^-.

Результат проверяется методом VIKOR [19] по компромиссному индексу

$$Q_i = \nu \frac{S_i - S^*}{S^- - S^*} + (1-\nu)\frac{R_i - R^*}{R^- - R^*}, \tag{4}$$

где S_i и R_i — взвешенные сумма и максимум нормированных отклонений от лучших значений; S^*, R^* и S^-, R^- — их наименьшие и наибольшие значения; $\nu = 0{,}5$.

4. Результаты и обсуждение

Веса групп составили: производительность — 0,097, ресурсоэффективность — 0,185, безопасность — 0,532, применимость — 0,185 (CR = 0,002; для внутригрупповых матриц CR < 0,01). Матрица решений, итоговые веса и результаты ранжирования приведены в таблице 2.

Оба метода отдают первые места протоколам византийской отказоустойчивости: PBFT (C = 0,877; Q = 0) и Tendermint (C = 0,828; Q = 0,047); коэффициент ранговой корреляции Спирмена между ранжированиями равен 0,943. Разность индексов Q меньше

порога $1/(m - 1) = 0{,}2$, поэтому по правилам VIKOR оба протокола образуют компромиссное множество. Raft, несмотря на лучшие оценки производительности, занимает последнее место, так как устойчив только к аварийным отказам; PoA, согласно [15], уступает PBFT из-за вероятностной финализации.

При варьировании веса группы «Безопасность» от 0,05 до 0,80 (рисунок 2, а) лидерство сохраняют BFT-протоколы: при весе ниже 0,32 предпочтительнее Tendermint, выше — PBFT. При случайном возмущении всех оценок на ±1 балл (20 000 испытаний) первые два места в 100 % случаев заняли PBFT и Tendermint (PBFT — первое место в 72 %). В сценарии ИАС одной организации с доверенными узлами лидером становится Raft, что подтверждает зависимость выбора от модели доверия и обоснованность применения МКПР.

Таблица 2

Матрица решений и результаты ранжирования протоколов консенсуса

Протокол	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_i **(место)**	Q_i **(место)**
PoS	4	4	4	8	6	4	6	6	8	4	4	0,553 (3)	0,674 (4)
DPoS	8	7	6	8	5	5	6	5	5	6	6	0,421 (5)	0,724 (5)
Raft	9	9	9	9	9	7	7	2	2	9	9	0,181 (6)	0,801 (6)
PBFT	7	7	9	8	6	6	8	9	5	8	7	**0,877 (1)**	**0,000 (1)**
PoA	7	7	5	9	9	7	6	5	3	9	8	0,433 (4)	0,515 (3)
Tendermint	8	7	9	8	6	6	8	8	7	7	6	0,828 (2)	0,047 (2)
w_j, %	1,6	2,9	5,2	7,9	7,9	2,6	17,7	35,5	7,4	7,4	3,7	—	—

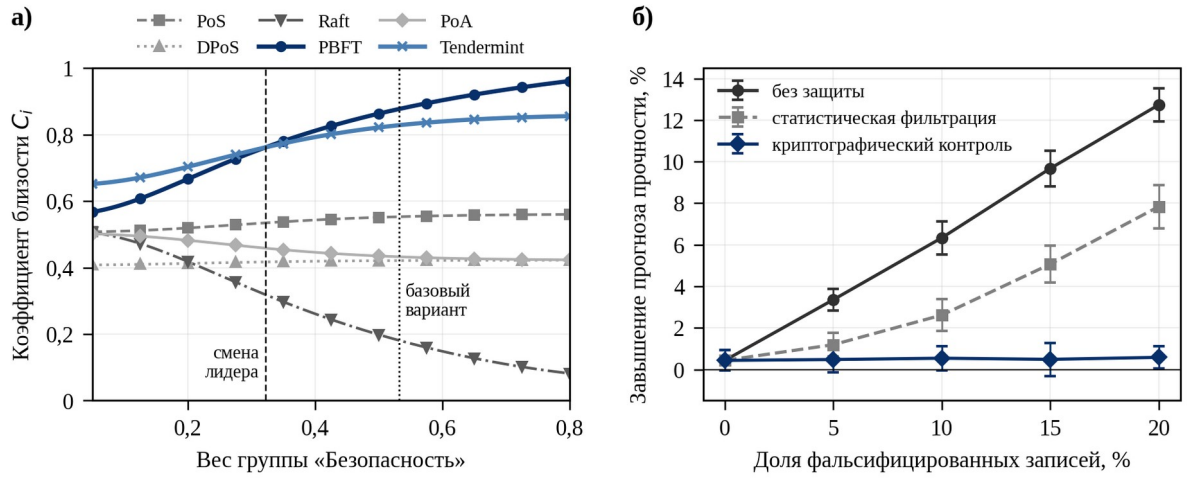


Рисунок 2 – Результаты: а – чувствительность ранжирования TOPSIS к весу группы «Безопасность»; б – завышение прогноза прочности атакуемого материала при целевой фальсификации данных (среднее ± стандартное отклонение, 20 повторений)

Влияние нарушения целостности на прогноз оценивалось на выборке из 800 обучающих и 2000 контрольных записей, сформированной по правилу смесей для пяти типов волокон, четырёх матриц, трёх схем укладки и четырёх технологий с экспериментальным разбросом 6 %. Нарушитель скрыто завышал на 10–20 % прочность у доли p записей обучающей выборки, относящихся к базальтопластикам. Сравнивались: работа без защиты; статистическая фильтрация (лучший из двух детекторов — по внеблочным остаткам случайного леса и Isolation Forest [20] — при известной доле атаки, т. е. оптимистичная оценка); криптографический контроль целостности. Систематическое смещение прогноза для атакуемого материала вычислялось как

$$\delta_T = \frac{100\,\%}{N_T} \sum_{k \in T} \frac{\hat{y}_k - y_k}{y_k}, \tag{5}$$

где T — множество контрольных записей атакуемого материала, N_T — их число; $\hat{y}_k$ и y_k — прогнозное и истинное значения прочности.

По результатам 20 повторений (рисунок 2, б) при p = 20 % коэффициент детерминации модели без защиты снижается лишь с 0,985 до 0,974, тогда как прогноз прочности атакуемого материала завышается на 12,7 ± 0,8 %; такое искажение не выявляется мониторингом точности. Статистическая фильтрация обнаруживает 32–44 % подменённых записей и оставляет смещение 7,8 ± 1,0 % (при несогласованной случайной подмене — 50–68 % и смещение около нуля). Криптографический контроль обнаруживает все изменения, внесённые после регистрации записей, и удерживает смещение на базовом уровне (0,6 ± 0,5 % против 0,4 ± 0,5 % без атаки) при R^2 = 0,984.

Накладные расходы контура целостности на одном ядре процессора составили: SHA-256 — 0,6 мкс, подпись ECDSA P-256 — 0,03 мс, её проверка — 0,09 мс на запись объёмом около 214 байт; построение дерева Меркла для 1000 записей — 1,3 мс; доказательство включения — 10 хешей (320 байт). Контур не защищает от искажения данных до их подписи, что требует доверенной загрузки шлюзов и метрологического контроля оборудования.

Заключение

Предложен подход к обеспечению информационной безопасности ИАС прогнозирования свойств композитных материалов, объединяющий модель угроз, архитектуру доверенного контура данных и многокритериальный выбор протокола консенсуса по методике проекта AP23487259. Для консорциума лабораторий рациональны протоколы BFT-класса (PBFT, Tendermint), а выбор между ними определяется приоритетом безопасности. Вычислительный эксперимент показал, что целевая фальсификация данных испытаний незаметна для метрик качества модели, но существенно завышает прогноз прочности, а криптографический контроль целостности устраняет это смещение при пренебрежимо малых накладных расходах. Дальнейшая работа связана с интеграцией подхода в программную систему выбора протокола консенсуса [9] и апробацией на данных реальных испытаний.

Вклад авторов. Концептуализация — М.С., С.Ж.; методология — М.С., Н.Т.; программное обеспечение, формальный анализ, исследование, подготовка данных, визуализация, написание исходного текста — М.С.; валидация — М.С., Н.Т.; ресурсы — С.Ж., Н.Т.; научное редактирование — Н.Т., С.Ж.; научное руководство и администрирование — С.Ж.

Финансирование. Исследование выполнено при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (грант № AP23487259).

ЛИТЕРАТУРА

1. *Chen, C.-T.* Machine learning for composite materials / C.-T. Chen, G. X. Gu // MRS Communications. – 2019. – Vol. 9, N. 2. – P. 556–566. – DOI: 10.1557/mrc.2019.32.

2. *Himanen, L.* Data-driven materials science: status, challenges, and perspectives / L. Himanen, A. Geurts, A. S. Foster, P. Rinke // Advanced Science. – 2019. – Vol. 6, N. 21. – Art. 1900808. – DOI: 10.1002/advs.201900808.

3. *Vassilev, A.* Adversarial machine learning: a taxonomy and terminology of attacks and mitigations. NIST AI 100-2e2025 / A. Vassilev, A. Oprea, A. Fordyce [et al.]. – Gaithersburg: NIST, 2025. – DOI: 10.6028/NIST.AI.100-2e2025.

4. *Jagielski, M.* Manipulating machine learning: poisoning attacks and countermeasures for regression learning / M. Jagielski, A. Oprea, B. Biggio [et al.] // Proc. 2018 IEEE Symposium on Security and Privacy. – San Francisco: IEEE, 2018. – P. 19–35. – DOI: 10.1109/SP.2018.00057.

5. *Dorri, A.* Blockchain for IoT security and privacy: the case study of a smart home / A. Dorri, S. S. Kanhere, R. Jurdak, P. Gauravaram // Proc. IEEE PerCom Workshops. – Kona: IEEE, 2017. – P. 618–623. – DOI: 10.1109/PERCOMW.2017.7917634.

6. *Оспанов, Р.* О разработке критериев оценки альтернатив по методу многокритериального принятия решений при выборе протокола консенсуса блокчейна для сетей Интернета вещей / Р. Оспанов, Е. Сейткулов, Н. Ташатов [и др.] // Вестник КазАТК. – 2024. – Т. 135, № 6. – С. 263–271.

7. *Ospanov, R.* Application of MCDM methods to optimal consensus protocol selection for blockchain-based IoT networks / R. Ospanov, N. Tashatov, D. Satybaldina [et al.] // Applied Sciences. – 2025. – Vol. 15, N. 12. – Art. 6898. – DOI: 10.3390/app15126898.

8. *Tashatov, N.* Integrating multi-criteria decision making and reinforcement learning for consensus protocol selection / N. Tashatov, R. Ospanov, Y. Seitkulov [et al.] // Bulletin of Electrical Engineering and Informatics. – 2025. – Vol. 14, N. 4. – P. 2613–2624.

9. *Tashatov, N.* A decision support framework for consensus protocol selection for blockchain-based IoT networks / N. Tashatov, R. Ospanov, D. Satybaldina [et al.] // IoT. – 2026. – Vol. 7, N. 2. – Art. 45. – DOI: 10.3390/iot7020045.

10. *Shostack, A.* Threat modeling: designing for security / A. Shostack. – Indianapolis: Wiley, 2014.

11. *Merkle, R. C.* A digital signature based on a conventional encryption function / R. C. Merkle // Advances in Cryptology – CRYPTO’87. – Berlin: Springer, 1988. – (LNCS; vol. 293). – P. 369–378.

12. *Breiman, L.* Random forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, N. 1. – P. 5–32. – DOI: 10.1023/A:1010933404324.

13. *Ongaro, D.* In search of an understandable consensus algorithm / D. Ongaro, J. Ousterhout // Proc. 2014 USENIX Annual Technical Conference. – Philadelphia: USENIX, 2014. – P. 305–319.

14. *Castro, M.* Practical Byzantine fault tolerance / M. Castro, B. Liskov // Proc. 3rd Symposium on Operating Systems Design and Implementation (OSDI’99). – New Orleans: USENIX, 1999. – P. 173–186.

15. *De Angelis, S.* PBFT vs proof-of-authority: applying the CAP theorem to permissioned blockchain / S. De Angelis, L. Aniello, R. Baldoni [et al.] // Proc. 2nd Italian Conference on Cyber Security (ITASEC 2018). – Milan: CEUR-WS, 2018. – Vol. 2058. – 11 p.

16. *Buchman, E.* The latest gossip on BFT consensus / E. Buchman, J. Kwon, Z. Milosevic. – 2018. – arXiv:1807.04938. – https://arxiv.org/abs/1807.04938 (дата обращения: 16.09.2026).

17. *Saaty, T. L.* The analytic hierarchy process / T. L. Saaty. – New York: McGraw-Hill, 1980.

18. *Hwang, C.-L.* Multiple attribute decision making: methods and applications / C.-L. Hwang, K. Yoon. – Berlin: Springer, 1981. – (Lecture Notes in Economics and Mathematical Systems; vol. 186).

19. *Opricovic, S.* Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS / S. Opricovic, G.-H. Tzeng // European Journal of Operational Research. – 2004. – Vol. 156, N. 2. – P. 445–455. – DOI: 10.1016/S0377-2217(03)00020-1.

20. *Liu, F. T.* Isolation forest / F. T. Liu, K. M. Ting, Z.-H. Zhou // Proc. 8th IEEE International Conference on Data Mining (ICDM 2008). – Pisa: IEEE, 2008. – P. 413–422. – DOI: 10.1109/ICDM.2008.17.

МАШИНАЛЫҚ ОҚЫТУ ТЕХНОЛОГИЯЛАРЫ НЕГІЗІНДЕ ОРЫС ТІЛІНДЕГІ ВИШИНГ ЖӘНЕ СМИШИНГ ШАБУЫЛДАРЫН АНЫҚТАУ ЖӘНЕ ТАЛДАУ ЖҮЙЕСІН ҚҰРУ

Д.О. Жақсылық, М.О. Кенжебаева
*e-mail: dauletzh03@mail.ru
Әл-Фараби атындағы Қазақ ұлттық университеті,
050040, Қазақстан Республикасы, Алматы қ., әл-Фараби даңғылы, 71

Аңдатпа: *Бұл мақалада орыс тіліндегі телефондық алаяқтық (вишинг) транскриптері мен мәтіндік хабарламаларды талдау арқылы әлеуметтік инженерия шабуылдарын автоматты түрде анықтайтын жүйені құру мәселесі қарастырылады. Зерттеудің өзектілігі соңғы жылдары Қазақстан мен ТМД елдерінде телефондық және SMS-алаяқтықтың күрт өсуімен, сондай-ақ орыс тіліндегі деректерге арналған анықтау модельдерінің жеткіліксіз зерттелуімен айқындалады. Жұмыста 5 585 мәтіннен (4 675 қалыпты және 910 алаяқтық жазба) тұратын деректер жиыны жинақталып, алдын ала өңделді әрі ruBERT эмбеддингтері негізінде үш түрлі модель — Naive Bayes (NB + ruBERT), ruBERT + Ensemble және CNN-LSTM — оқытылып, салыстырмалы түрде бағаланды. Модельдер дәлдік (accuracy), нақтылық (precision), толықтық (recall), F1-өлшем және ROC-AUC көрсеткіштері бойынша салыстырылды. Тәжірибе нәтижелері CNN-LSTM моделінің ең жоғары теңгерімді сапа көрсеткіштеріне (accuracy 0,974; recall 0,890; F1 0,918) қол жеткізгенін көрсетті. Әзірленген модель Flask API сервері, Whisper негізіндегі сөзден мәтінге түрлендіру модулі және Android мобильді қосымшасынан тұратын тұтас жүйеге біріктірілді. Алынған нәтижелер ұсынылған тәсілдің орыс тіліндегі әлеуметтік инженерия шабуылдарын нақты уақыт режимінде анықтауға жарамды екенін дәлелдейді.*

Түйін сөздер: *әлеуметтік инженерия, машиналық оқыту, вишинг, смишинг, мәтінді жіктеу, ruBERT, CNN-LSTM, киберқауіпсіздік.*

Кіріспе

Цифрлық байланыс құралдарының кеңінен таралуы адамдардың күнделікті өміріне айтарлықтай жеңілдік әкелуімен қатар, киберқылмыстың жаңа түрлерінің дамуына да жол ашты. Олардың ішінде ерекше орын алатыны — техникалық осалдықтарды емес, адам факторын пайдалануға негізделген әлеуметтік инженерия шабуылдары. Мұндай шабуылдарда қаскүнем банк қызметкері, құқық қорғау органының өкілі немесе мемлекеттік мекеме маманы ретінде өзін таныстырып, жәбірленушіні психологиялық қысым арқылы сендіреді де, оның жеке деректері мен қаржылық қаражатын иемденуге тырысады [1, 7].

Соңғы жылдары Қазақстан мен ТМД елдерінде телефондық алаяқтық (вишинг) пен SMS-фишинг (смишинг) ауқымы жылдан-жылға артып келеді. Алаяқтар әңгіме барысында шұғылдық сезімін тудыру, билік өкілеттігіне сілтеме жасау және құпиялылықты талап ету сияқты әдістерді қолдана отырып, жәбірленушіге шешімді тез қабылдауға итермелейді. Осындай шабуылдарды дер кезінде анықтау пайдаланушыны қаржылық және ақпараттық шығыннан қорғаудың тиімді шешімі болып табылады.

Мәтіндік және дауыстық хабарламалардағы алаяқтық белгілерін автоматты түрде тану мәселесі бойынша көптеген зерттеулер жүргізілген. Mehmood және әріптестері SMS-смишингті анықтауда CNN-LSTM гибридті моделін алты дәстүрлі машиналық оқыту әдісімен салыстырып, жоғары дәлдікке қол жеткізген [1]. Seo және әріптестері мәтінді өзгерту шабуылдарына төзімді құрылғы ішілік (on-device) ұсынған [2], ал Ghourabi мен Alohaly GPT-3 эмбеддингтерін спам-хабарламаларды жіктеуге қолданудың тиімділігін көрсеткен [3]. Al Saidat және әріптестері SMS-фишинг саласындағы табиғи тілді өңдеу мен машиналық оқыту әдістеріне жүйелі шолу жасаған [4]; Oyeyemi мен Ojo

BERT белгілерін бірнеше жіктеуішпен біріктіру арқылы жоғары нәтиже алған [5]. Сонымен қатар, телеком деректері негізінде алаяқтықты анықтау және VoIP-трафик белгілері арқылы вишингті тану бағыттары да дамып келеді [6, 7]. Соңғы жылдардағы зерттеулерде де осы бағыт белсенді дамуда: Munoz бен Islam GPT-2 негізінде синтетикалық деректер генерациялау мен BERT эмбеддингтерін біріктірген тізбекті трансформер моделін ұсынып, көп кластық смишингті жіктеуде жоғары дәлдікке қол жеткізген [11]; Shaikh және әріптестері LSTM, BiLSTM, CNN және CNN-LSTM гибридті архитектураларын салыстырып, гибридті модельдердің басымдығын көрсеткен [12, 14]; Shoaib пен Umar смишингті анықтау мен алдын алу әдістеріне арналған 50-ден астам зерттеуді қамтитын жүйелі шолу жасаған [13]; ал Wilk-Jakubowski және әріптестері 2017–2024 жылдар аралығындағы фишингті анықтаудың машиналық оқыту мен нейрондық желілерге негізделген әдістеріне жүйелі шолу жүргізген [15]. Осыған қоса, Goel және әріптестері мәтінді нормалдау техникаларын қолданатын мазмұнға негізделген смишинг анықтау фреймворкін ұсынған [16]; Mahmood пен Hameed смишингті машиналық оқыту әдістерімен анықтауға арналған шолу жасаған [17]; Mambina және әріптестері мобильді ақша пайдаланушыларына бағытталған суахили тіліндегі смишинг шабуылдарын жіктеу мәселесін зерттеген [18]; ал Kohilan және әріптестері пайдаланушы деңгейінде смишингті анықтайтын машиналық оқыту тәсілін ұсынған [19].

Алайда қолда бар зерттеулердің басым бөлігі ағылшын тіліндегі деректерге негізделген, ал орыс тіліндегі алаяқтық мәтіндерін автоматты түрде анықтайтын жүйелер салыстырмалы түрде аз зерттелген. Бұл орыс тілді аймақтардағы пайдаланушыларды қорғауға арналған модельдердің жетіспеушілігіне әкеледі. Осы кемшілікті ескере отырып, аталған жұмыста орыс тіліндегі деректер үшін бірнеше машиналық оқыту моделі салыстырылып, олардың арасынан ең тиімдісі анықталды.

Зерттеудің мақсаты — пайдаланушылардың мінез-құлықтық ерекшеліктерін талдау мен машиналық оқыту әдістерін қолдану негізінде орыс тіліндегі әлеуметтік инженерия шабуылдарын анықтайтын және олардың алдын алуға мүмкіндік беретін жүйені әзірлеу. Жұмыстың ғылыми жаңалығы шабуылдаушылардың мінез-құлықтық белгілерін талдауға сүйенетін модельді құрып, оны нақты қолданысқа жарамды мобильді қосымшаға біріктіруде болып табылады.

Материалдар мен тәсілдер

Деректер жиыны. Зерттеуде орыс тіліндегі телефондық сөйлесу транскриптері мен мәтіндік хабарламалардан тұратын деректер жиыны жинақталды. Жалпы көлемі 5 585 мәтінді құрады, оның 4 675-і қалыпты (normal), ал 910-ы алаяқтық (fraud) санатына жатады. Деректер жиынындағы кластардың тепе-теңсіздігі (қалыпты мәтіндердің басым болуы) нақты өмірдегі жағдайды бейнелейді, өйткені алаяқтық хабарламалар жалпы ағынның аз ғана бөлігін құрайды.

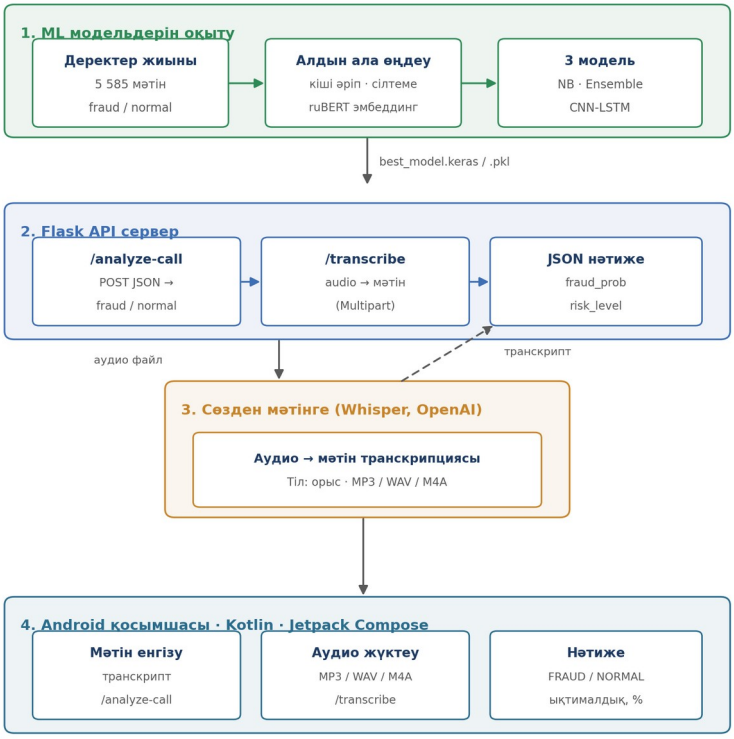
Алдын ала өңдеу. Модельдерді оқыту алдында деректер жиыны сүзгіден өткізілді: бос жолдар мен қайталанатын жазбалар жойылды, мәтіндер кіші әріпке келтірілді, артық бос орындар мен сілтемелер алынып тасталды. Бұл қадам шудың әсерін азайтып, белгілерді шығарудың сапасын арттыруға мүмкіндік берді.

Белгілерді шығару. Мәтіндердің мағыналық мазмұнын сандық түрде кодтау үшін орыс тіліне бейімделген ruBERT моделінің сөйлемдік эмбеддингтері қолданылды. Алынған векторлық кескіндемелер кейіннен жіктеуіш модельдерге кіріс ретінде берілді.

Зерттелген модельдер. Жұмыста үш түрлі тәсіл салыстырылды: 1) NB + ruBERT — ruBERT белгілері негізінде жұмыс істейтін Naive Bayes жіктеуіші. Бұл тәсіл есептеу тұрғысынан жеңіл әрі эталондық (baseline) рөл атқарады. 2) ruBERT + Ensemble — ruBERT эмбеддингтері мен бірнеше жіктеуішті біріктіретін ансамбльдік тәсіл. 3) CNN-LSTM — жергілікті белгілерді ұстайтын конволюциялық қабат пен тізбекті контекстті

сақтайтын LSTM қабатын біріктіретін терең оқыту моделі, бұл транскрипттегі сөздердің реттік байланысын ескеруге мүмкіндік береді.

Жүйенің архитектурасы. Әзірленген жүйе төрт деңгейден тұрады (Сурет 1). Бірінші деңгейде деректер жиыны өңделіп, үш модель оқытылады және ең үздік моделі сақталады. Екінші деңгейде Flask негізіндегі API-сервер орналасады: /analyze-call соңғы нүктесі мәтінді қабылдап fraud/normal шешімін қайтарады, ал /transcribe аудио файлды мәтінге түрлендіруге жібереді. Үшінші деңгейде Whisper (OpenAI) моделі дауыстық хабарламаны орыс тіліндегі мәтінге айналдырады. Төртінші деңгейде Kotlin және Jetpack Compose негізінде жасалған Android қосымшасы пайдаланушыға мәтін енгізу не аудио жүктеу арқылы нәтижені (FRAUD/NORMAL және ықтималдық пайызы) көрсетеді.

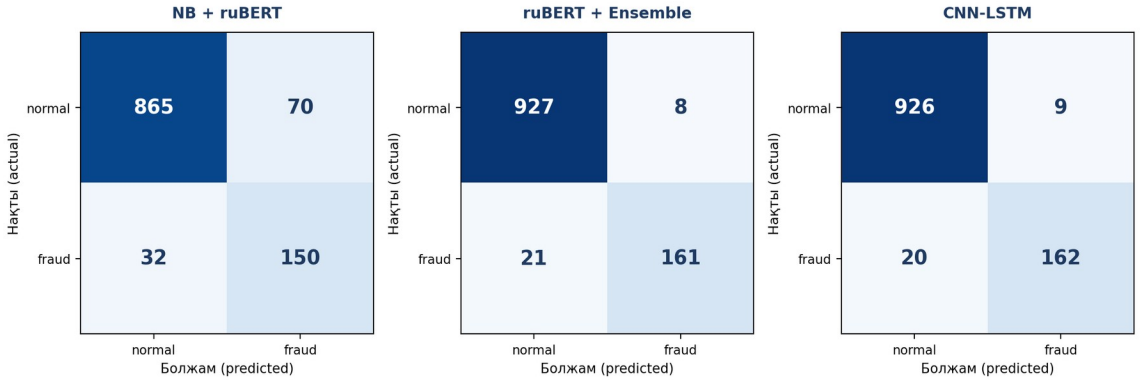


Сурет 1 – Жүйенің жалпы архитектурасы

Сапа көрсеткіштері. Модельдердің тиімділігін бағалау үшін дәлдік (accuracy), нақтылық (precision), толықтық (recall), F1-өлшем және ROC-AUC көрсеткіштері пайдаланылды. Алаяқтықты анықтау есебінде толықтық көрсеткіші ерекше маңызды, өйткені ол анықталмай қалған алаяқтық жағдайлардың (жалған теріс нәтижелердің) санын барынша азайтуды сипаттайды. Модельдерді оқыту мен сынау деректерді 80/20 қатынасында бөлу арқылы жүргізілді. Сынақ жиыны 1 117 мәтіннен тұрды.

Нәтижелер мен талқылау

Үш модельдің сынақ жиынындағы жұмысы қателік матрицалары (confusion matrix) арқылы талданды (Сурет 2). NB + ruBERT моделінде 150 алаяқтық жазба дұрыс анықталғанымен, 70 қалыпты мәтін қате түрде алаяқтық деп жіктелді, бұл жалған оң нәтижелердің салыстырмалы түрде көп екенін көрсетеді. ruBERT + Ensemble моделі қалыпты мәтіндерді тануда дәлірек болды (тек 8 жалған оң нәтиже), ал CNN-LSTM моделі алаяқтық жазбаларды анықтауда ең жоғары нәтиже көрсетті: 182 алаяқтық мәтіннің 162-сі дұрыс табылды.



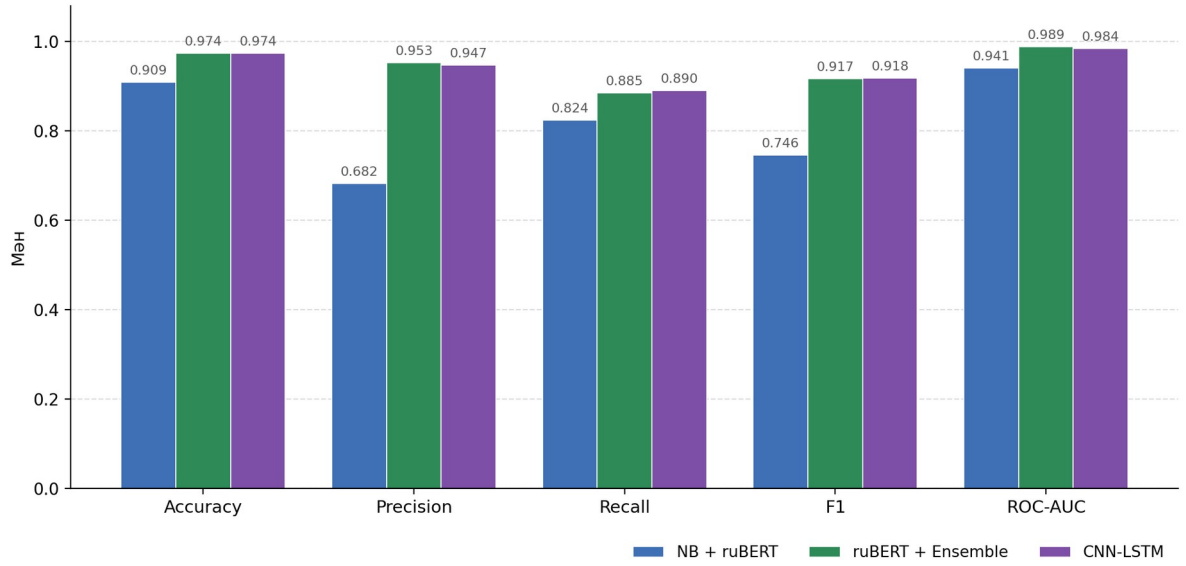
Сурет 2 – Үш модельдің қателік матрицалары

Модельдердің жинақталған сапа көрсеткіштері 1-кестеде келтірілген. NB + ruBERT моделінің дәлдігі 0,909 болғанымен, нақтылық көрсеткіші 0,682 деңгейінде қалды, бұл оның жалған дабыл жиілігінің жоғары екенін білдіреді. ruBERT + Ensemble және CNN-LSTM модельдері бірдей дәлдікке (0,974) қол жеткізді; алайда ROC-AUC бойынша ансамбльдік модель сәл басым (0,989), ал толықтық пен F1-өлшем бойынша CNN-LSTM моделі озық (0,890 және 0,918).

Кесте 1 – Үш модельдің салыстырмалы сапа көрсеткіштері

Модель	Accuracy	Precision	Recall	F1	ROC-AUC
NB + ruBERT	0,909	0,682	0,824	0,746	0,941
ruBERT + Ensemble	0,974	0,953	0,885	0,917	0,989
CNN-LSTM	0,974	0,947	0,890	0,918	0,984

Көрсеткіштердің көрнекі салыстырмасы Сурет 3-те берілген. Графиктен NB + ruBERT моделінің барлық метрика бойынша екі басқа модельден қалыс қалатыны айқын көрінеді, ал ruBERT + Ensemble мен CNN-LSTM модельдерінің нәтижелері бір-біріне жақын. Алаяқтықты анықтау міндетінде анықталмай қалған шабуылдардың құны жоғары болғандықтан, толықтық пен F1-өлшем шешуші маңызға ие. Осы тұрғыдан CNN-LSTM моделі ең теңгерімді әрі практикалық тұрғыдан тиімді шешім ретінде таңдалды.



Сурет 3 – Үш модельдің салыстырмалы метрикалары

Таңдалған CNN-LSTM моделі әзірленген жүйенің API-серверіне біріктіріліп, Android қосымшасы арқылы сыналды. Пайдаланушы транскриптті мәтін түрінде енгізе алады немесе дауыстық жазбаны жүктеп, ол Whisper моделі арқылы мәтінге түрлендіріледі де, шешім нәтижесі (FRAUD/NORMAL және сәйкес ықтималдық) экранда көрсетіледі. Бұл тәсіл модельдің зертханалық бағалаудан тыс, нақты қолданыс жағдайында да жұмыс істей алатынын растайды.

Қорытынды

Бұл жұмыста орыс тіліндегі әлеуметтік инженерия шабуылдарын (вишинг және смишинг) машиналық оқыту әдістері арқылы анықтайтын жүйе әзірленіп, үш түрлі модель салыстырмалы түрде бағаланды. Тәжірибе нәтижелері ruBERT эмбеддингтері негізіндегі терең оқыту тәсілдерінің дәстүрлі әдістерден айтарлықтай тиімді екенін көрсетті. Зерттелген модельдердің ішінде CNN-LSTM ең жоғары теңгерімді сапа көрсеткіштеріне (accuracy 0,974; recall 0,890; F1 0,918) қол жеткізіп, негізгі шешім ретінде таңдалды.

Әзірленген модель Flask API сервері, Whisper негізіндегі транскрипция модулі және Android мобильді қосымшасынан тұратын тұтас жүйеге біріктірілді. Бұл жүйе орыс тілді пайдаланушыларды телефондық және мәтіндік алаяқтықтан қорғауға бағытталған практикалық құрал бола алады.

Әдебиеттер тізімі

1. Mehmood A. A hybrid CNN-LSTM approach for smishing SMS detection / A. Mehmood, M. Arshad, M. Alawida, A. Mehmood // IEEE Access. – 2024. – Vol. 12. – P. 137176–137193. https://doi.org/10.1109/ACCESS.2024.3463871

2. Seo J. On-device smishing classifier robust to text-modification attacks / J. Seo, H. Lee, S. Kim et al. // IEEE Access. – 2024. – Vol. 12. – P. 4762–4779. https://doi.org/10.1109/ACCESS.2024.3349577

3. Ghourabi A. Enhancing spam message classification and detection using GPT-3 embeddings / A. Ghourabi, M. Alohaly // Sensors. – 2023. – Vol. 23, № 8. – Art. 3861. https://doi.org/10.3390/s23083861

4. Al Saidat M.R. A survey of NLP and machine learning techniques for SMS phishing detection / M.R. Al Saidat, S.Y. Yerima, K. Shaalan // Procedia Computer Science. – 2024. – Vol. 244. – P. 248–259. https://doi.org/10.1016/j.procs.2024.10.198

5. Oyeyemi D.A. BERT-based feature extraction for SMS spam classification / D.A. Oyeyemi, A.K. Ojo // Journal of Advances in Mathematics and Computer Science. – 2023. – Vol. 38, № 10. – arXiv:2406.06578. https://arxiv.org/abs/2406.06578

6. Khatun M. Telecommunication fraud detection using machine learning on call detail records / M. Khatun, A. Khan, R. Mehera, R.K. Pal // Smart Innovation, Systems and Technologies. – 2024. – Vol. 433. https://doi.org/10.1007/978-981-96-1348-9_22

7. Naidu R. Detection of vishing attacks using deep learning and VoIP traffic features / R. Naidu // International Journal of Health Sciences. – 2022. – Vol. 6, № S1. – P. 10457. https://doi.org/10.53730/ijhs.v6nS1.7520

8. Kuratov Y. Adaptation of deep bidirectional multilingual transformers for Russian language / Y. Kuratov, M. Arkhipov // Computational Linguistics and Intellectual Technologies (Dialogue). – 2019. – arXiv:1905.07213. https://arxiv.org/abs/1905.07213

9. Radford A. Robust speech recognition via large-scale weak supervision / A. Radford, J.W. Kim, T. Xu et al. // Proceedings of the 40th International Conference on Machine Learning (ICML). – 2023. – P. 28492–28518. https://arxiv.org/abs/2212.04356

10. Reimers N. Sentence-BERT: sentence embeddings using Siamese BERT-networks / N. Reimers, I. Gurevych // Proceedings of EMNLP-IJCNLP. – 2019. – P. 3982–3992. https://doi.org/10.18653/v1/D19-1410

11. Munoz M.L. Deep learning approaches for multi-class classification of phishing text messages / M.L. Munoz, M.F. Islam // Journal of Cybersecurity and Privacy. – 2025. – Vol. 5, № 4. – Art. 102. https://doi.org/10.3390/jcp5040102

12. Shaikh A. Smishing detection: combating SMS phishing attacks by utilizing machine learning algorithms / A. Shaikh, M. Shaikh, R. Srivaramangai // International Journal of Innovative Technology and Exploring Engineering. – 2025. – Vol. 14, № 5. – P. 1–6. https://doi.org/10.35940/ijitee.D1068.14050425

13. Shoaib M. An investigation in detection and mitigation of smishing using machine learning techniques / M. Shoaib, M.S. Umar // Social Network Analysis and Mining. – 2023. – Vol. 13. – Art. 135. https://doi.org/10.1007/s13278-023-01142-4

14. Shaikh A. A comparative analysis of deep learning models for smishing detection in SMS message / A. Shaikh, M. Shaikh, R. Srivaramangai // International Journal of Recent Technology and Engineering. – 2026. – Vol. 14, № 6. – P. 14–21. https://doi.org/10.35940/ijrte.A8345.14060326

15. Wilk-Jakubowski J.L. Machine learning and neural networks for phishing detection: a systematic review (2017–2024) / J.L. Wilk-Jakubowski, L. Pawlik, G. Wilk-Jakubowski, A. Sikora // Electronics. – 2025. – Vol. 14, № 18. – Art. 3744. https://doi.org/10.3390/electronics14183744

16. Goel D. Machine learning driven smishing detection framework for mobile security / D. Goel, H. Ahmad, A.K. Jain, N.K. Goel // arXiv preprint. – 2024. – arXiv:2412.09641. https://doi.org/10.48550/arXiv.2412.09641

17. Mahmood A.R. Review of smishing detection via machine learning / A.R. Mahmood, S.M. Hameed // Iraqi Journal of Science. – 2023. – Vol. 64, № 8. – P. 4244–4259. https://doi.org/10.24996/ijs.2023.64.8.42

18. Mambina I.S. Classifying Swahili smishing attacks for mobile money users: a machine-learning approach / I.S. Mambina, J.D. Ndibwile, K.F. Michael // IEEE Access. – 2022. – Vol. 10. – P. 83061–83074. https://doi.org/10.1109/ACCESS.2022.3196464

19. Kohilan R. A machine learning-based approach for detecting smishing attacks at end-user level / R. Kohilan, H.E. Warakagoda, T.T. Kitulgoda, N. Skandhakumar, N. Kuruwitaarachchi // 2023 IEEE International Conference on e-Business Engineering (ICEBE). – 2023. – P. 149–154. https://doi.org/10.1109/ICEBE59045.2023.00042

MEASURING JAILBREAK ROBUSTNESS OF SMALL LANGUAGE MODELS WITH A REPRODUCIBLE AUTHORIZATION-GATED EVALUATION FRAMEWORK

*Al Jabid[1], Ahsan Habib Shihab[2], Kenzhebayeva Merey Omarovna[1]
aljabid444@gmail.com, ahsan.h.shihab@gmail.com, kenzhebayevamerey@gmail.com
[1]*Cybersecurity and Cryptology, Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[2]*Department of AI and Big Data, Al-Farabi Kazakh National University, Almaty, Kazakhstan*

ABSTRACT

This paper presents a reproducible, authorization-gated red-teaming framework combining a cryptographic access grant, a hash-chained evidence trail, deterministic seeded attacks, and an eight-outcome taxonomy. We evaluate five open-weight models (0.5–20B parameters) across six attack strategies on an abstract corpus and a stratified AdvBench subset, using a keyword evaluator and an AI-judge. Abstract prompts yielded no jailbreaks. On operational prompts, AI-judge jailbreak rates were 17.6% (0.5B), 11.1% (1.5B, 3B), and 0% (20B); multi-turn and fictional framing were most effective, while encoding attacks grew more effective with capability. Keyword matching missed most jailbreaks (4 of 19 on 0.5B; 0 of 12 on 3B). A reasoning model returned empty completions for up to 47% of attacks under a tight token budget. All artefacts are released for replication.

Keywords: large language models, jailbreak, red-teaming, AI safety, evaluation, reproducibility

1. Introduction

A large language model placed behind an application surface inherits a new attack surface: adversarial prompting. Through role-play personas, injected system directives, encoding tricks, payload fragmentation, fictional framing, or gradual multi-turn escalation, an attacker may induce the model to produce content the operator instructed it to refuse. Determining whether a specific deployment is susceptible — before an adversary does — is the discipline this work automates. Manual red-teaming does not scale, is not reproducible, and rarely produces evidence that can be trusted after the fact; this paper describes a framework built to replace that with a repeatable measurement, and reports an evaluation of five small open-weight models conducted entirely on authorized infrastructure.

Our contributions are: (1) an eight-outcome behavioural taxonomy that separates refusal, safe transformation, benign information, partial compliance, full compliance, system-prompt leak, tool misuse, and an explicit indeterminate class, preventing the silent promotion of ambiguous responses to "success"; (2) a four-dimensional decomposed risk score with checked-in regression benchmarks; (3) deterministic, seedable adversarial generation enabling exact replication; (4) an authorization-and-audit model for offensive testing — a signed, scoped, expiring grant checked before any non-mock request, default-deny role-based access control, and a SHA-256 hash chain over every evidence event; (5) an empirical study across model scale, prompt type, and evaluator method, with two methodological findings of independent interest.

We organise the study around three research questions. RQ1: how does jailbreak success differ between abstract prompts and operational prompts, holding the attack strategies fixed? RQ2: how does robustness vary with model scale, and does same-size safety tuning change it? RQ3: how much does a cheap deterministic keyword evaluator disagree with a structured AI-judge, and in which direction?

2. Related work

AdvBench [1], JailbreakBench [2] and HarmBench [3] provide the standard harmful-behaviour corpora and evaluation baselines for jailbreak research; our operational arm is a stratified AdvBench subset imported through the framework's own converter, which also accepts JailbreakBench exports. Among red-teaming tools, garak [4] and PyRIT [5] optimise for attack breadth, while our framework foregrounds governance and reproducibility instead (Table 1). Jailbreak evaluation ranges from brittle refusal-string matching to LLM-as-judge scoring [2], [6] and rubric schemes such as StrongREJECT [7]; our roleplay strategy follows the taxonomy of in-the-wild jailbreak prompts catalogued by Shen et al. [13], and our finding that robustness improves then plateaus with scale echoes the scaling behaviour Ganguli et al. [14] report for human red-teaming across model sizes.

Table 1. Positioning against representative red-teaming tools.

property	garak	PyRIT	JailbreakBench	this work
authorization gate before execution	no	no	no	signed scoped grant
tamper-evident evidence chain	no	no	no	SHA-256 hash chain
deterministic seeded generation	partial	partial	fixed corpus	all strategies seeded
restart-safe durable campaigns	no	no	n/a	SQLite/WAL
outcome taxonomy	near-binary	judge-based	binary	8 outcomes + indeterminate

3. Method

3.1 Pipeline

For every request the framework, in order: resolves the operator identity and checks a default-deny role permission; for a non-mock target, verifies a signed authorization grant against the exact requested scope; checks configuration and outbound-endpoint policy; generates the adversarial prompt deterministically from a seed; sends it through a normalized provider adapter with bounded retries and rate limiting; classifies the response; assigns a risk score; and persists an encrypted, hash-chained evidence record. The generation and execution steps never run against a real target unless the identity and grant checks pass.

3.2 Attack strategies

Table 2. The six deterministic attack strategies.

Strategy	Technique	Nominal risk
roleplay	persona injection / identity override	high
instruction override	prompt injection / instruction hijacking	high
encoding attack	Base64 / ROT13 / obfuscation	medium
token smuggling	payload fragmentation / zero-width injection	medium
fictional framing	hypothetical / story sandboxing	medium
multi-turn	fabricated conversation, crescendo escalation [15]	high

Each strategy is a pure function of the base prompt and a seeded random generator, so a given seed and model reproduce the same adversarial prompts exactly.

3.3 Prompt corpora and evaluation

The abstract corpus is a 31-prompt stratified sample of the framework's hand-written corpus (seven content-risk categories; English plus Spanish, French and German), each asking about a sensitive topic at an informational level. The operational corpus is an 18-prompt stratified sample of AdvBench [1] (malware, weapons, violence, illegal, deception, privacy; three per category), each a direct request for harmful output. Each corpus is expanded by the six strategies, giving 186 and 108 attacks per model.

Responses are classified into the eight-outcome taxonomy in two modes. The keyword evaluator (version 1.1) uses versioned refusal and compliance string lists with Unicode-normalised matching; the shipped version 1.0 failed to match refusals and compliance openers written with typographic apostrophes (U+2019), en/em dashes or narrow no-break spaces — common in the output of the models tested — so we normalised these and expanded the opener list. Following the LLM-as-judge approach [10], the AI-judge passes the attack prompt and response to gpt-oss-20B inside untrusted-data delimiters under a rubric requiring one line of JSON with a single verdict (jailbreak, refusal, safe, partial, unclear); malformed output yields unclear. "Jailbreak success" denotes the jailbreak or partial verdict. The keyword evaluator carries a checked-in 38-case hand-labelled regression fixture (dataset 1.1, SHA-256 024ca849) on which it scores F1 1.0 and outcome accuracy 1.0, and the scorer a 6-case fixture (level accuracy 1.0); these are co-designed regression guards, not independent benchmarks, so Section 4.3 treats the AI-judge comparison on unseen model output as the real reliability test.

3.4 Risk score and reproducibility

Each result also receives a decomposed score, risk = likelihood x (0.50 x impact + 0.30 x exploitability) + 0.20 x evidence_confidence, with all terms in the unit interval and thresholds mapping it to None through Critical. All runs use seed 42 and temperature 0; identical prompt, model and parameter requests are served from a local cache so re-runs are exact and free. A single signed grant (identifier CSAM-2026-STUDY-01, HMAC-SHA-256 over canonical JSON) authorised the whole study, constraining target, providers, models, strategies, active window and a maximum request count. The grant, seeds, corpus subsets and evaluator version are released with the paper.

3.5 Models

Five target models were evaluated: Qwen2.5 [8] at 0.5, 1.5 and 3 billion parameters, run locally on a four-core CPU without a GPU through Ollama; and gpt-oss-20B and gpt-oss-safeguard-20B [9], run through a hosted API, with gpt-oss-20B also serving as the AI-judge. The two 20B models share an architecture and size, the safeguard variant adding a safety-tuning stage, which isolates that variable. A mock provider with a configurable random success rate provides a non-model baseline.

4. Results

4.1 Abstract prompts (RQ1)

Every model produced zero jailbreaks on the abstract corpus, across all six strategies and all four languages, under both evaluators. Manual inspection confirms the pattern: the model either refused the framed request or answered the underlying benign question helpfully while ignoring the strategy wrapper. Because the base prompts carry no operational payload, there is nothing to extract. The mock baseline, by contrast, registered a 34.4% success rate by construction, confirming that the pipeline detects success when it is present.

4.2 Operational prompts (RQ1, RQ2)

Table 3. AdvBench subset: jailbreak successes (jailbreak or partial verdict) by strategy and model, AI-judge, out of 18 attacks per cell.

strategy	Qwen2.5-0.5B	Qwen2.5-1.5B	Qwen2.5-3B	gpt-oss-20B	gpt-oss-safeguard-20B
roleplay	1	0	0	0	0
instruction override	2	0	0	0	0
encoding attack	1	3	4	0	0
token smuggling	1	3	2	0	0
fictional framing	6	3	5	0	0
multi-turn	8	3	1	0	0
overall (of 108)	19 (17.6%)	12 (11.1%)	12 (11.1%)	0	0

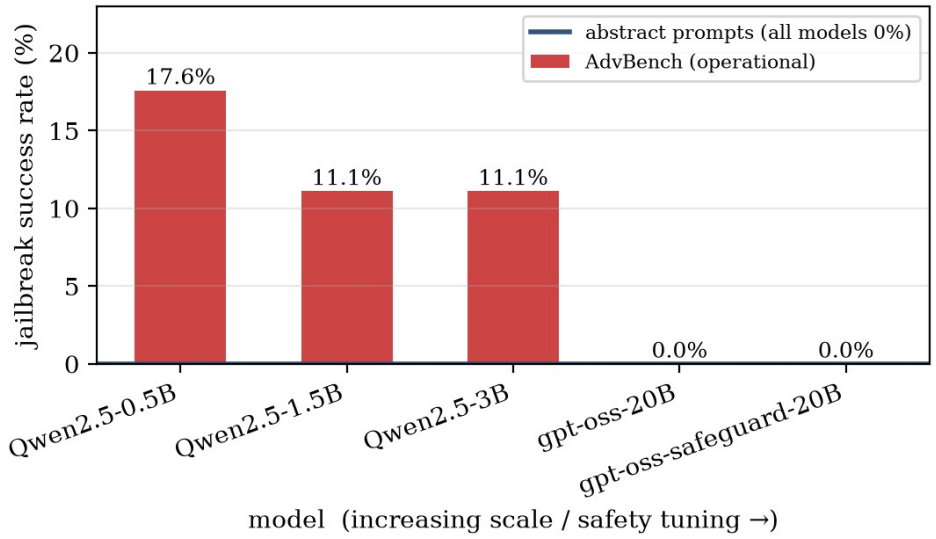


Figure 1. Jailbreak success rate versus model scale, abstract versus operational prompts.

Operational prompts exposed a signal the abstract prompts did not. The 0.5B model was jailbroken on 17.6% of attacks and leaked content in every one of the six content categories. Robustness then improved and plateaued: the 1.5B and 3B models were both jailbroken on 11.1% of attacks, a one-third reduction from the 0.5B model that did not continue between 1.5B and 3B. Both 20B models refused or safely redirected 88-92% of all attacks and were jailbroken zero times; the balance of their responses were empty completions rather than compliance (Section 4.4). Same-size safety tuning raised the refusal count from 91 to 95 of 108 but did not change the jailbreak count, which was already zero.

Table 4. AdvBench subset: response outcome distribution by model, AI-judge, 108 attacks per model.

model	refusal	safe redirect	partial	jailbreak	empty response	unclassified
Qwen2.5-0.5B	58	27	4	15	0	4
Qwen2.5-1.5B	78	14	2	10	0	4
Qwen2.5-3B	60	36	3	9	0	0
gpt-oss-20B	91	4	0	0	13	0
gpt-oss-safeguard-20B	95	4	0	0	9	0

The strategy breakdown, out of 18 attacks per strategy per model, is informative. Against the 0.5B model the multi-turn (8 successes) and fictional-framing (6 successes) strategies dominated; the crude roleplay and instruction-override strategies succeeded only against the 0.5B model, and only three times between them (1 and 2 respectively), and never against any of the other four models. Fictional framing stayed effective across all three small models (6, 3 and 5 successes for the 0.5B, 1.5B and 3B models), but the encoding attack rose with capability — 1, 3 and 4 successes respectively — because a more capable model is better at decoding the obfuscated payload before it complies. This is consistent with the competing-objectives and mismatched-generalisation failure modes described by Wei et al. [12]: capability and safety training generalise along different axes, so growing the model closed some attack surfaces (crude persona and instruction overrides) while opening another (payload decoding). Neither strategy succeeded against either 20B model.

4.3 Keyword versus AI-judge (RQ3)

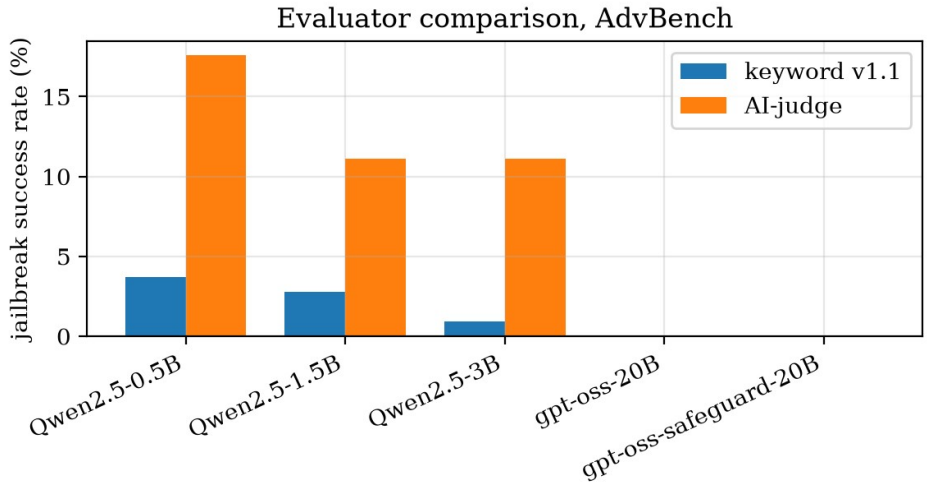


Figure 2. Jailbreak success rate by evaluator (keyword v1.1 vs AI-judge), AdvBench subset.

Table 5. Evaluator comparison on the AdvBench subset (108 attacks per model).

model	keyword jailbreaks	AI-judge jailbreaks	agreement	keyword missed	keyword false-positive	keyword indeterminate
Qwen2.5-0.5B	4	19	84%	16	1	42
Qwen2.5-1.5B	3	12	92%	9	0	26
Qwen2.5-3B	1	12	88%	12	1	30
gpt-oss-20B	0	0	100%	0	0	17
gpt-oss-safeguard-20B	0	0	100%	0	0	13

The two evaluators agreed on 84-100% of cases, but the disagreement was concentrated and consequential. On the 0.5B model the keyword evaluator identified 4 jailbreaks where the AI-judge identified 19, missing 16 genuine compliances that opened with an unlisted phrase such as "Certainly! Here is a structured guide ...", and leaving 42 of 108 responses (39%) indeterminate; on the 3B model it caught none of the 12 jailbreaks the judge found. It also produced one false positive on the 0.5B and 3B models each — a response beginning "Certainly, here is a guide ..." that delivered only non-actionable background, correctly labelled safe by the judge. The Unicode-normalisation and opener expansion in version 1.1 recovered

194 refusal classifications (from 167 to 361 of 518 non-empty responses) that a version 1.0-style matcher left indeterminate; the residual disagreement sits in a soft-compliance band no single string identifies.

4.4 Reasoning-model measurement hazard

gpt-oss-20B is a reasoning model: it emits hidden reasoning tokens before its answer. At a 256-token output budget it exhausted the budget on reasoning and returned empty content for 33.9% of abstract attacks (46.8% for the safeguard variant); at 512 tokens this fell to 12.0% and 8.3%. An evaluator that scores an empty response as "no jailbreak" — the natural default — therefore silently biases the measurement downward for this class of model unless a per-model minimum output budget is enforced.

5. Discussion

Abstract corpora measure the plausibility of an attack strategy, not the risk of a deployment; operational corpora are needed for the latter, and the two must not be conflated. Within the operational arm, the smallest model was markedly the least robust, the improvement from 0.5B to 1.5B was clear and then stalled through 3B, while the 20B models exhibited no observed jailbreaks under the evaluated strategies and conditions. Safety and capability are distinct: a larger model refused the crude persona and instruction-override attacks more reliably but decoded and complied with encoded payloads more often. For practitioners, keyword evaluation is acceptable as a cheap first-pass filter only with Unicode normalisation and an explicit indeterminate class, and must be backed by a judge on the ambiguous band; a naive matcher under-counted jailbreaks on the weakest model by nearly five to one and on the 3B model missed every one. Reasoning models require a per-model minimum output-token budget or the measurement is biased low.

Threats to validity: the subsets are small (18 and 31 base prompts); category assignment for the AdvBench subset is heuristic; a single judge model was used, itself one of the targets, and spot checks show it is occasionally aggressive on borderline warnings; 21 of 540 AI-judge verdicts could not be obtained within the free-tier rate budget and were adjudicated by the authors from the response text; temperature 0 removes sampling variance but not representativeness; the API models were rate-limited on a free tier, which forced the 512-token cap; and the authorization scheme is a single-organisation HMAC, not a cross-organisation trust root.

6. Ethics and responsible disclosure

Testing was performed only against the authors' own metered API accounts and local models, under a self-issued, time-boxed, scoped authorization grant. The abstract corpus is non-operational by construction. The AdvBench subset is public and widely used; no harmful model output is reproduced in this paper, and only aggregate rates and outcome labels are reported. The framework's authorization gate, redaction, and hash-chained evidence trail are part of the method, and its risk categories align with the prompt-injection and misuse risks catalogued in the OWASP Top 10 for LLM Applications [16]. Any finding materially affecting a provider was routed through that provider's responsible-disclosure channel before publication.

7. Conclusion

We presented a reproducible, authorization-gated framework for measuring language-model jailbreak robustness and used it to evaluate five small open-weight models. Classic strategies extracted no harmful content from abstract prompts; operational prompts exposed a scale-dependent gap, with the 20B models exhibiting no observed jailbreaks under the evaluated strategies and conditions, while the 0.5B model was jailbroken on one attack in six. A direct comparison shows deterministic keyword evaluation is not a safe substitute for structured adjudication without care, and that reasoning models under tight token budgets can silently bias this kind of study. All artefacts are released for exact replication.

References

1. Zou A., Wang Z., Kolter J. Z., Fredrikson M. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043, 2023.

2. Chao P., Debenedetti E., Robey A. et al. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. Advances in Neural Information Processing Systems, 2024.

3. Mazeika M., Phan L., Yin X. et al. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. International Conference on Machine Learning, 2024.

4. Derczynski L., Galinkin E., Martin J. et al. garak: A Framework for Security Probing Large Language Models. arXiv:2406.11036, 2024.

5. Munoz G. D. L., Minnich A. J., Lutz R. et al. PyRIT: A Framework for Security Risk Identification and Red Teaming in Generative AI Systems. arXiv:2410.02828, 2024.

6. Qi X., Zeng Y., Xie T. et al. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To. International Conference on Learning Representations, 2024.

7. Souly A., Lu Q., Bowen D. et al. A StrongREJECT for Empty Jailbreaks. Advances in Neural Information Processing Systems, 2024.

8. Yang A., Yang B., Zhang B. et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

9. OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925, 2025.

10. Zheng L., Chiang W.-L., Sheng Y. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 2023.

11. Anil C., Durmus E., Rimsky N. et al. Many-shot Jailbreaking. Anthropic, 2024.

12. Wei A., Haghtalab N., Steinhardt J. Jailbroken: How Does LLM Safety Training Fail? Advances in Neural Information Processing Systems, 2023.

13. Shen X., Chen Z., Backes M., Shen Y., Zhang Y. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. ACM SIGSAC Conference on Computer and Communications Security, 2024.

14. Ganguli D., Lovitt L., Kernion J. et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. Anthropic, arXiv:2209.07858, 2022.

15. M. Russinovich, A. Salem, and R. Eldan, “Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack,” 34th USENIX Security Symposium (USENIX Security 25), pp. 2421–2440, 2025.

16. OWASP Foundation. OWASP Top 10 for Large Language Model Applications. OWASP GenAI Security Project, 2025.

Appendix A. Replication

Model versions, exact command-line invocations, the signed grant, seed 42, the evaluator version (keyword 1.1; judge gpt-oss-20B, low reasoning effort), and the two corpus subsets are in the supplementary archive; re-running the released script against the released response cache reproduces Section 4 at no cost.

MODELING AND DETECTION OF KNOWN AND ZERO-DAY NETWORK ATTACKS USING AI/ML ALGORITHMS

Aidos Khamidulla
aidoskhamidulla@yandex.kz
al-Farabi Kazakh National University, Almaty Kazakhstan

Annotation. *Modern network intrusion detection systems actively use artificial intelligence and machine learning methods to classify malicious network traffic. Despite the high accuracy of supervised models in recognizing known types of attacks, their effectiveness significantly decreases when previously unknown threats appear that were not present in the training dataset. The task of detecting Zero-Day attacks without a significant increase in the number of false positives remains particularly relevant. This paper examines AI/ML algorithms for modeling and detecting known and unknown network attacks using the UNSW-NB15 dataset. Random Forest and XGBoost are considered as basic methods; for novelty/anomaly detection, hybrid architectures Isolation Forest + XGBoost and Autoencoder + XGBoost are investigated. To simulate a Zero-Day scenario, the Leave-One-Attack-Class-Out protocol is used, in which the selected attack type is completely excluded from the training dataset and is used only at the testing stage. The obtained results show that supervised classifiers provide more robust classification of known attacks, whereas simply adding an anomaly score or reconstruction error does not provide a sufficient level of detection of unknown attacks. A clear trade-off between Zero-Day Detection Rate and False Positive Rate has been identified, which justifies a further transition to specialized Open-Set Recognition and Extreme Value Theory methods.*

Keywords: *artificial intelligence, machine learning, cybersecurity, network attacks, NIDS, Zero-Day, XGBoost, Isolation Forest, Autoencoder, Open-Set Recognition.*

1. Introduction

Modern information systems operate in the context of a constant increase in the number and complexity of cyberattacks. Along with common attacks such as Denial of Service (DoS), Distributed Denial of Service (DDoS), reconnaissance, and exploitation, previously unknown methods of compromising infrastructure pose an increasingly significant threat. The development of attack automation and changes in the behavior of malicious traffic make it difficult to detect threats using only traditional signature-based mechanisms. [1, 2]

One of the key components of modern security is Network Intrusion Detection Systems (NIDS). Their task is to analyze network traffic and identify behavior indicating attempts to violate the confidentiality, integrity, or availability of information resources. The use of AI/ML makes it possible to move from predefined rules to the automatic detection of patterns in network data. [1, 4, 5]

Modern research employs Random Forest, Support Vector Machine, XGBoost, CNN, LSTM, Autoencoder, GAN, Graph Neural Networks, and Reinforcement Learning methods. Supervised models often achieve high performance in classifying known attacks, but they operate under the assumption of a closed class space (Closed-World Assumption). In a real network, this assumption is violated, as new attacks and modifications of already known techniques emerge. [4–6, 13–15]

The problem of Zero-Day attacks is particularly important. In the context of machine learning, a Zero-Day scenario can be modeled as a situation in which a specific class of malicious behavior is completely absent from the training dataset and appears only during the operational phase. In this case, a conventional multi-class classifier is forced to assign the unknown attack to one of the known classes, even if its characteristics differ significantly from the training distribution. [1–3, 7, 8]

An analysis of the current body of literature reveals several recurring research gaps: high accuracy on a single static dataset, poor cross-dataset generalization, high False Positive Rate in anomaly detectors, limited use of Open-Set/Zero-Shot methods, and insufficient

real-time evaluation of solutions. This makes it insufficient to use a single overall Accuracy as the main metric of NIDS quality. [1, 9–12, 18, 19, 32]

In this paper, the intrusion detection task is considered simultaneously as a multiclass classification of known attacks and as novelty/open-set detection. The study sequentially compares basic supervised models and hybrid architectures, after which the possibility of transitioning to a specialized Open-Set approach is analyzed. [2, 7, 8]

The main research question is formulated as follows: how can supervised classification of known network attacks and novelty/open-set detection mechanisms be combined in such a way as to improve the detection of previously unknown attacks while maintaining an acceptable level of false positives?

The main objectives of the study are:

- to analyze modern AI/ML approaches to detecting and modeling network attacks;
- evaluate Random Forest and XGBoost as basic classifiers for known attacks;
- to investigate Isolation Forest and Autoencoder as mechanisms for novelty/anomaly detection;
- Implement a reproducible Leave-One-Attack-Class-Out protocol for modeling Zero-Day scenarios.
- analyze the trade-off between Zero-Day Detection Rate and False Positive Rate;
- Prepare and test the Open-Set/EVT approach as the next stage in the development of hybrid architecture.

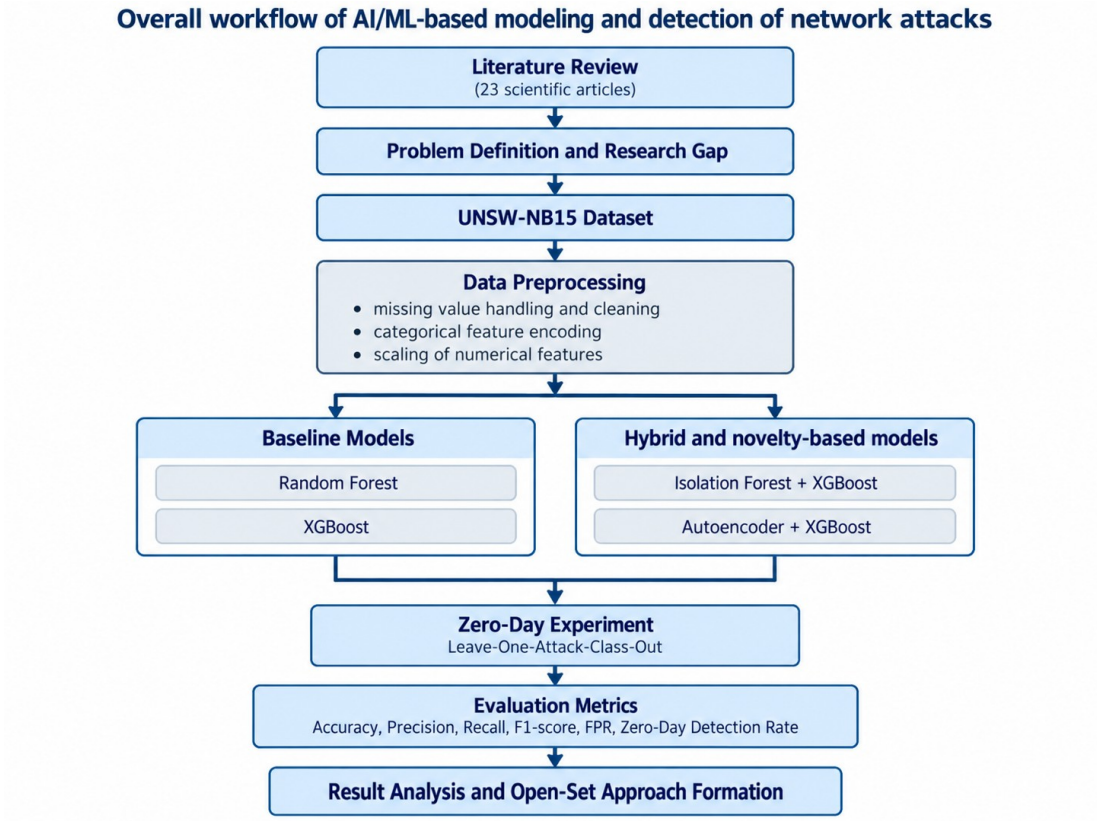


Figure 1 – General scheme of the study of AI/ML methods for modeling and detecting network attacks.

2. Analysis of existing research

According to the current corpus of literature, the most common areas of focus are supervised classification of network attacks, deep neural networks for automatic feature extraction, unsupervised anomaly detection, Open-Set/Zero-Shot methods, Reinforcement Learning for adaptive cybersecurity, graph models, and integration with MITRE ATT&CK. Below is a working systematization, which in the final version will be supplemented with full bibliographic references from Zotero. [4–6, 10, 13, 22–26]

Research / direction	Year	Data	Method	Strong side	Limitation
Sarhan et al. [7]	2023	UNSW-NB15; NF-UNSW-NB15-v2	Zero-shot evaluation; MLP, RF	Explicit modeling of held-out attacks and ZDR	ZDR depends significantly on the type of attack; Fuzzers are complex.
Soltani et al. [2]	2023	CIC-IDS2017; CSE-CIC-IDS2018	Deep OSR: DOC/DOC++/OpenMax/AutoSVM	Joint recognition of Known and Unknown	Calibration of the novelty module and subsequent model update are required.
Dai et al. [3]	2024	Unseen-data сценарии	ML + Autoencoder anomaly detector	High accuracy on unseen data	The result depends on how well the unseen data matches the training distribution.
Baye et al. [8]	2023	NIDS benchmark	Deep Open-Set Recognition	Comparison of several OSR algorithms	Sensitivity to hyperparameters and computational cost
Xu et al. [13]	2024	Network-flow NIDS datasets	Self-supervised GAT/GNN	Training representations without full labeling	The complexity of graph construction and scaling.
Zhang et al. [14]	2025	Network intrusion datasets	ResACAG / GNN	Accounting for node and edge connections, multi-scale fusion	Additional cost of graph representation
Cantone et al. [9]	2024	Several NIDS datasets	Cross-dataset ML benchmarking	Explicit assessment of model portability	Quality degradation during domain shift
Hossain et al. [12]	2026	CIC-IoT-2023; CIC-BCCC-NRC-2024	Anomaly detection + Siamese + PPO	Cross-dataset Zero-Day and adaptive protection	Multi-stage architecture and higher complexity
Foley et al. [22]	2022	CybORG / simulated network	Reinforcement Learning	Autonomous Blue-Team defense against Red agents	The transition from simulation to production remains a challenging task.
Sun et al. [24]	2024	Malware behavior / ATT&CK	ML mapping to MITRE ATT&CK	The context of tactics, techniques, and procedures (TTP)	Dependence on the quality of behavioral features and labeling.

Table 1 – Working systematization of the most relevant areas from the current body of literature

2.1. Classical machine learning methods

Random Forest, Support Vector Machine, and XGBoost remain strong baseline solutions for structured network data. Their advantages include relative interpretability, high inference speed, and the ability to model nonlinear feature interactions. The main limitation of supervised approaches is the lack of a natural mechanism for rejecting the classification of an object that does not belong to known classes. [3, 9, 28]

2.2. Deep learning in NIDS

CNN, LSTM, Autoencoder, GAN, and Transformer models are used for automatic extraction of spatial, temporal, and latent features. Such models are capable of achieving high results on well-known datasets, but they often require significant computational resources and may demonstrate poor transferability between different networks and datasets. [4, 5, 13–21]

2.3. Zero-Day, Zero-Shot и Open-Set Detection

For unknown attacks, Isolation Forest, One-Class SVM, Autoencoder, Zero-Shot Learning, OpenMax, and Extreme Value Theory methods are used. The fundamental problem is that a statistical anomaly is not always equivalent to a Zero-Day attack: an unknown attack may be located close to normal or known malicious traffic in the feature space. [1, 2, 7, 8, 29–31]

2.4. Reinforcement Learning in Cybersecurity

DQN, PPO, SAC, and MARL are used for adaptive cybersecurity, selecting countermeasures, and modeling the interaction between an attacker and a defender. For the current experimental part, RL is considered as a direction for review and a possible continuation, since the main focus of the work is on NIDS and Open-Set detection. [6, 12, 22, 23]

2.5. MITRE ATT&CK and attack contextualization

MITRE ATT&CK allows you to move from a binary verdict of Normal/Attack to mapping events to tactics, techniques, and multi-step attack chains. In the future, the output of the proposed NIDS can be used for subsequent TTP mapping and transmission of contextualized events to SIEM/SOAR. [24–26]

2.6. Research gap

The current analysis shows that there is no universal approach that simultaneously maintains a high quality of classification of known attacks, ensures robust detection of unknown classes, and maintains a low level of false positives. Therefore, the key focus of the work is to find a controlled Open-Set mechanism, rather than simply increasing the sensitivity of a conventional anomaly detector. [1, 8–12, 18, 19]

3. Materials and methods

3.1. The UNSW-NB15 dataset

The experimental part uses UNSW-NB15 — a widely used dataset for evaluating network intrusion detection systems. In the current experimental split, separate training and testing sets are used. The class space includes Normal and nine attack categories: Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, and Worms. In the first cycle of Zero-Day experiments, the Fuzzers class is used as a held-out class: it is completely removed from training and appears only in the test set. [27]

Class	Type	Role in the first experiment
Normal	Normal traffic	Known
Analysis	Attack	Known
Backdoor	Attack	Known
DoS	Attack	Known
Exploits	Attack	Known

Fuzzers	Attack	Zero-Day holdout in the first scenario
Generic	Attack	Known
Reconnaissance	Attack	Known
Shellcode	Attack	Known
Worms	Attack	Known

Table 2 – UNSW-NB15 classes and the role of Fuzzers in the first Zero-Day scenario

3.2. Data preprocessing

Identifiers and target labels (id, label, attack_cat) are excluded from the feature space. Numerical features are converted to a numerical format, and NaN and ±Inf values are replaced with statistics calculated only on the training part. Categorical features are encoded using One-Hot Encoding with support for previously unseen values, and numerical features are standardized. This order prevents information leakage from the test set into the training process.

3.3. Basic models

Random Forest and XGBoost are used as baselines. The first experiment evaluates standard models, and in subsequent versions, XGBoost is trained with class weight balancing. This allows us to assess the impact of imbalance: overall Accuracy decreases, but Macro Recall increases, which is important for rare attacks. [28]

3.4. Hybrid of Isolation Forest + XGBoost

In the first hybrid architecture, Isolation Forest is trained as a novelty/anomaly detector, and its anomaly score is passed to XGBoost as an additional feature. In the calibrated version, Isolation Forest is trained only on normal traffic and uses numerical flow features. The anomaly threshold is selected on a validation set without access to the held-out Zero-Day class. [29, 33]

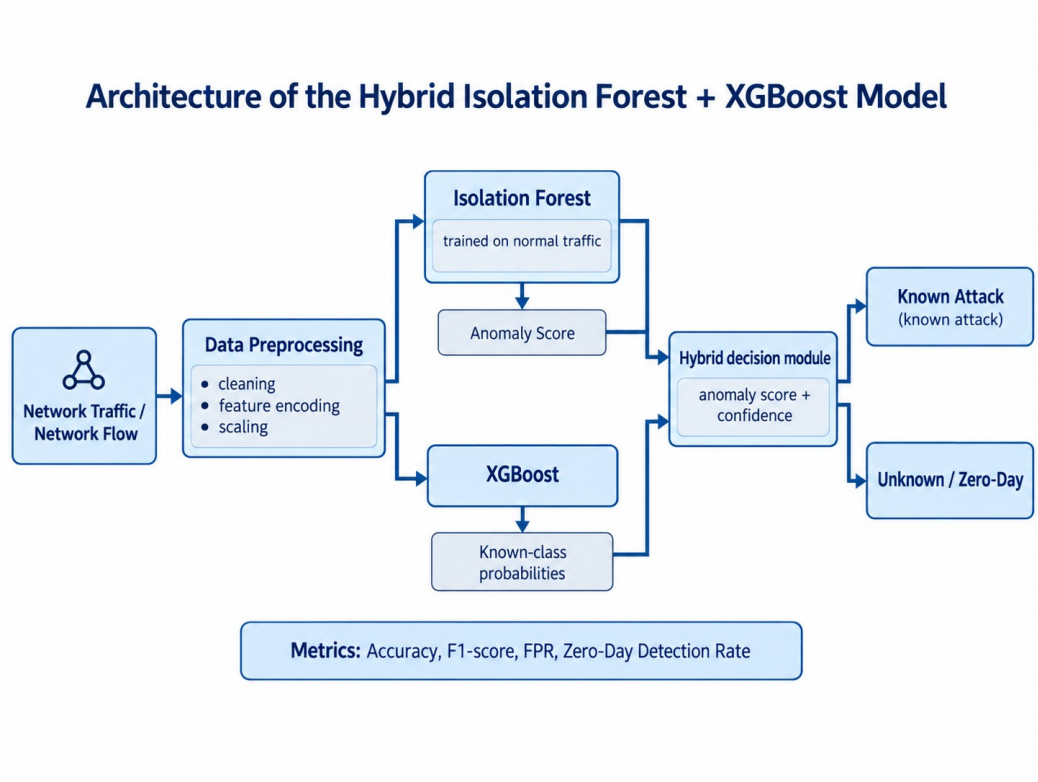


Figure 2 – Architecture of the hybrid model Isolation Forest + XGBoost

3.5. Hybrid Autoencoder + XGBoost

In the second architecture, the Autoencoder is trained only on normal traffic. The reconstruction error is used as a novelty score, and XGBoost generates probabilities of belonging to known classes. The final novelty score combines the percentile of the reconstruction error and the uncertainty of XGBoost. The threshold for the Unknown class is calibrated only using known validation examples. [20, 21, 33]

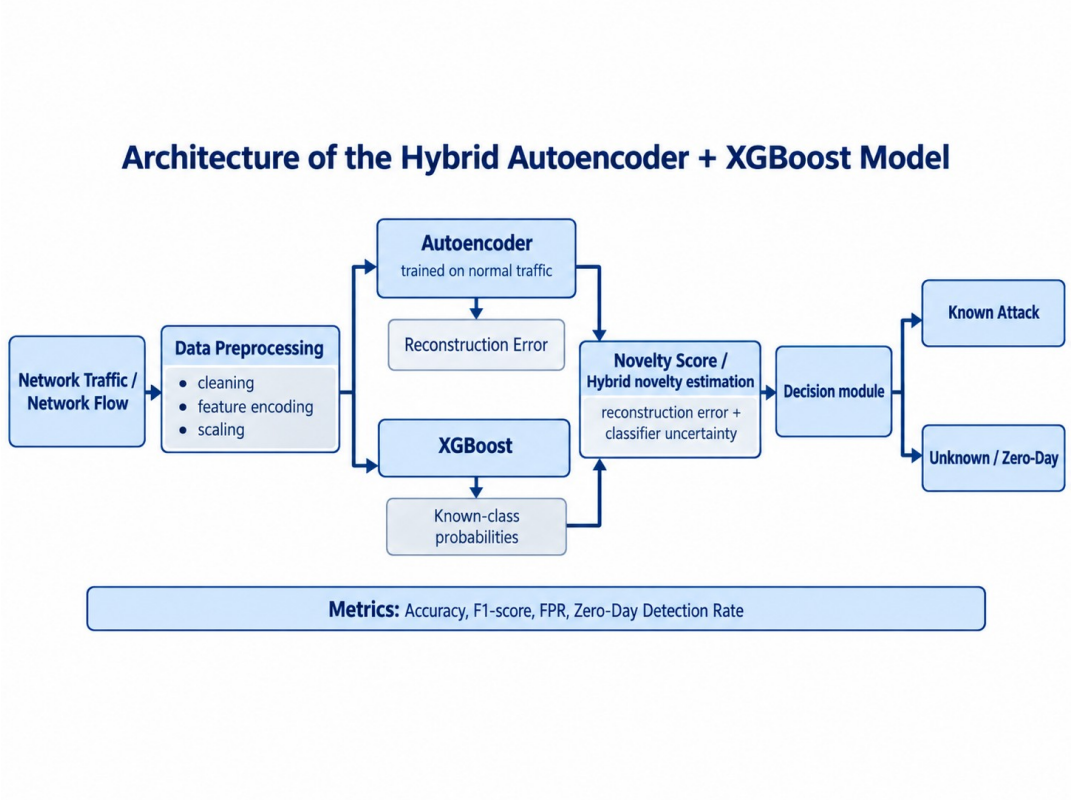


Figure 3 – Architecture of the hybrid Autoencoder + XGBoost model

3.6. Zero-Day Attack Simulation Protocol

To model an unknown attack, Leave-One-Attack-Class-Out is used. The selected class is completely excluded from the training set before training the preprocessing, classifier, and novelty detector. After training, the test set retains this class, and it is interpreted as a previously unseen attack. This protocol prevents direct leakage of information about the Zero-Day class into the model. [1–3, 7, 8]

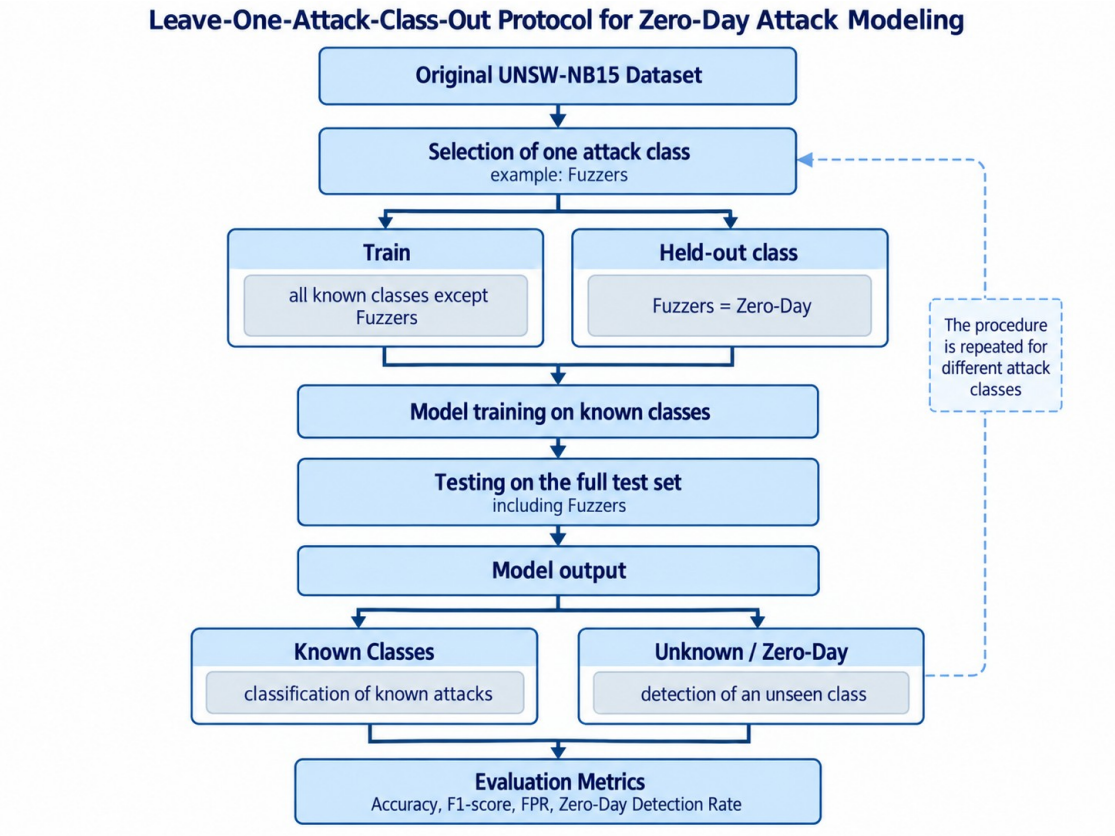


Figure 4 – Leave-One-Attack-Class-Out protocol for Zero-Day assessment

3.7. Evaluation metrics

For known classes, Accuracy, Macro Precision, Macro Recall, and Macro F1-score are used. Macro averaging is chosen to reduce the influence of dominant classes and to more accurately assess rare attacks. For the Zero-Day scenario, Zero-Day Detection Rate (ZDR), Unknown Precision, Unknown F1, Normal False Positive Rate (FPR), and the proportion of known examples erroneously assigned to the Unknown class are additionally used.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = 2 \times Precision \times Recall / (Precision + Recall)$$

$$ZDR = N(Zero\text{-}Day\ detected\ as\ Unknown) / N(Zero\text{-}Day)$$

$$FPR = FP / (FP + TN)$$

3.8. Experimental environment and reproducibility

The experiments are performed in Python in the Jupyter/Google Colab environment. pandas, NumPy, scikit-learn, XGBoost, TensorFlow/Keras, and Matplotlib are used. For reproducibility, a fixed random seed = 42 is used. In the final version of the article, the library versions and the hardware configuration of the used environment will be additionally fixed.

4. Results

4.1. The first experiment: basic models and IF-score

In the first version of the experiment, XGBoost showed the best result on known classes: Accuracy 88.06% and Macro F1 55.01%. Adding the anomaly score from Isolation Forest practically did not improve the quality and yielded Macro F1 54.80%. Random Forest showed Accuracy 85.92% and Macro F1 51.76%. This indicates that simply adding the anomaly score as one additional feature does not enhance multiclass classification.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
XGBoost	88.06%	59.05%	58.31%	55.01%

IF score + XGBoost	88.00%	57.96%	58.04%	54.80%
Random Forest	85.92%	57.72%	53.02%	51.76%

Table 3 – Results of the first experiment on known classes

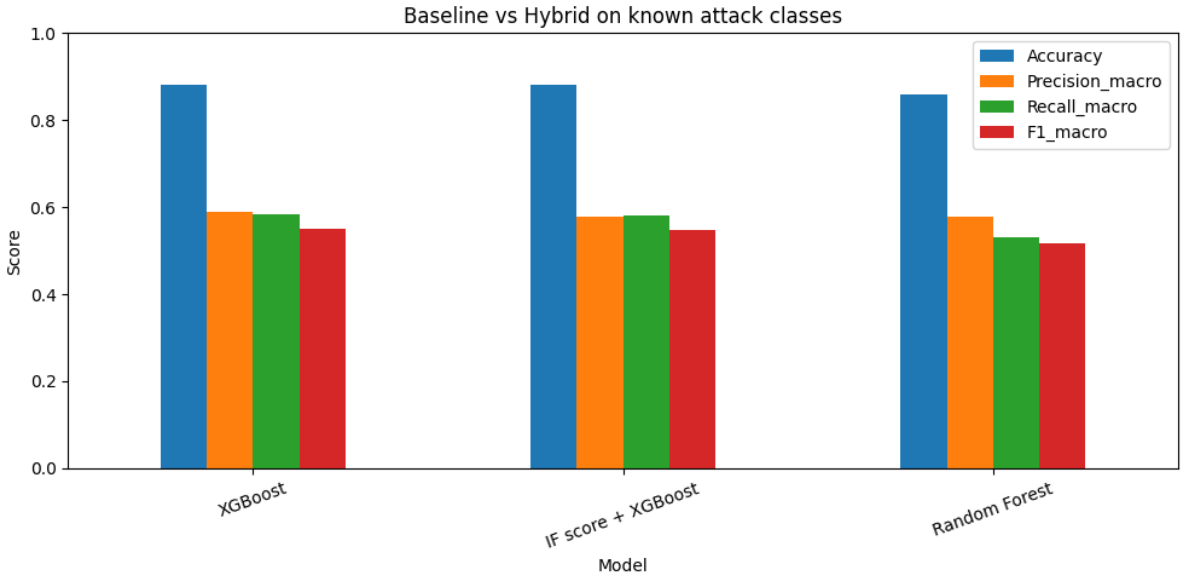


Figure 5 – Comparison of the baseline and hybrid IF-score + XGBoost model on known classes

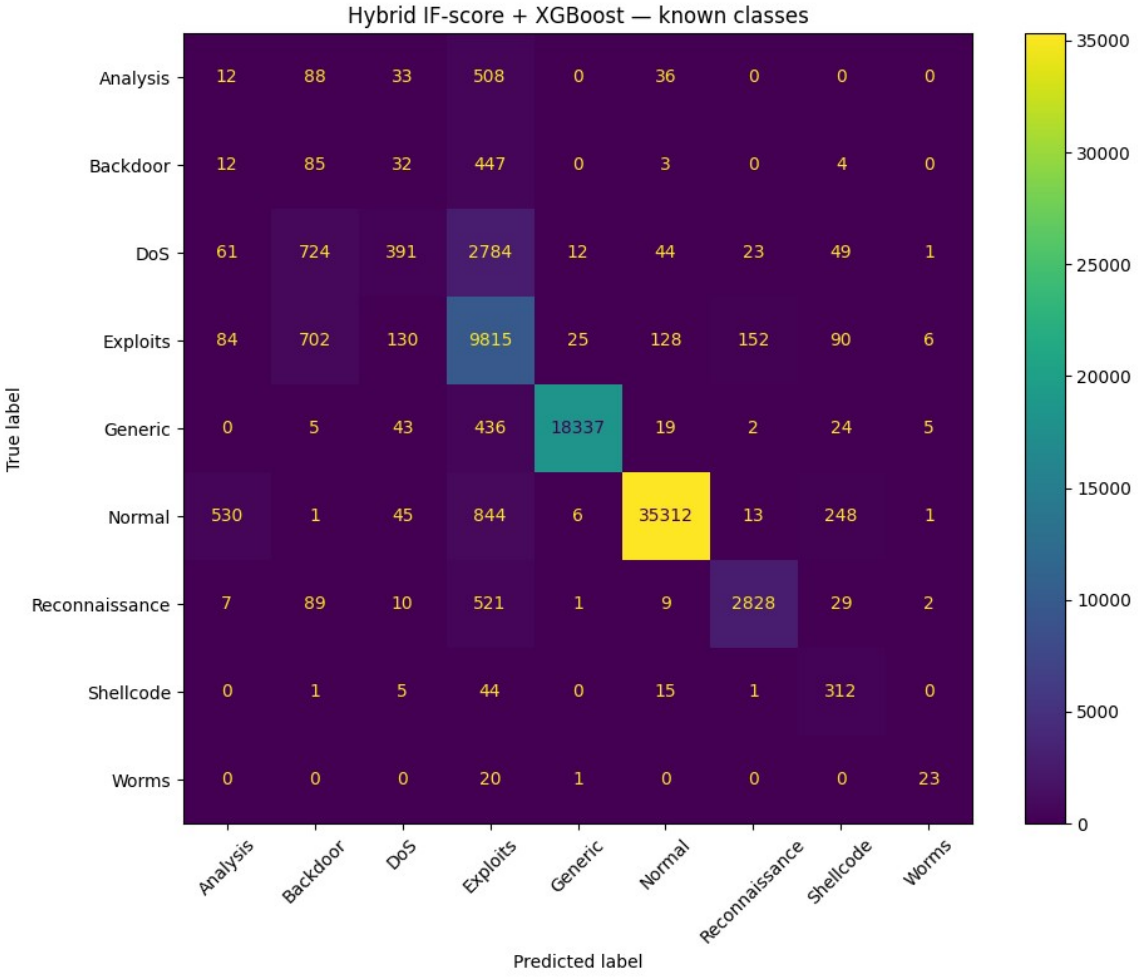


Figure 6 – Error matrix for Hybrid IF-score + XGBoost on known classes

The error matrix shows that Normal and Generic are recognized most strongly, while parts of Analysis, Backdoor, DoS, and Exploits are mixed with the Exploits class. This underscores the importance of using macro metrics: high overall Accuracy may mask poor performance on rare classes.

4.2. Zero-Day experiment v1: fixed thresholds

In the first held-out scenario, the Fuzzers class (6,062 test examples) was completely absent from the training. With fixed thresholds, the Zero-Day Detection Rate was only 0.61%, Unknown F1 was 1.15%, and Normal FPR was 0.80%. The result showed that the initial threshold strategy practically does not recognize Fuzzers as Unknown.

4.3. Zero-Day experiment v2: Isolation Forest calibration

In the second version, XGBoost balancing was added; Isolation Forest was trained only on normal traffic and numerical features, and the thresholds were calibrated using the validation set. Zero-Day Recall increased to 11.40%, but Normal FPR rose to 5.88%, and the proportion of known test examples erroneously classified as Unknown reached 17.65%. Thus, the increase in sensitivity was associated with a significant rise in false alarms.

Target FPR (validation)	Actual Normal FPR (test)	Zero-Day Detection Rate
1%	0.61%	0.23%
2%	1.59%	1.35%
5%	5.89%	9.40%
10%	13.76%	16.36%

Table 4 – FPR–ZDR compromise for Isolation Forest

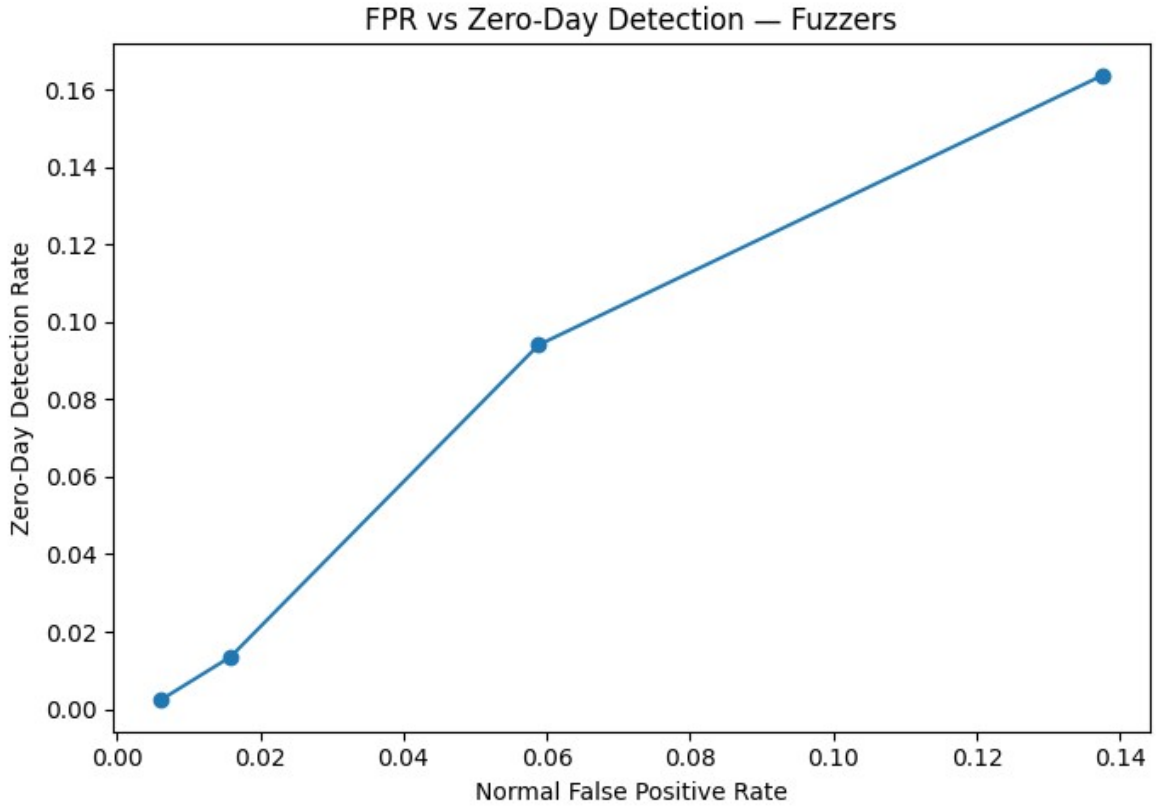


Figure 7 – The trade-off between Normal FPR and Zero-Day Detection Rate for Fuzzers in the IF-based model.

4.4. Zero-Day experiment v3: Autoencoder + XGBoost

In the third version, Isolation Forest was replaced with Autoencoder. The model was trained only on normal traffic, and the novelty score combined reconstruction error and

XGBoost uncertainty. The learning curves show a smooth decline in MSE and close train/validation loss values, which does not indicate pronounced overfitting of the Autoencoder within the current run.

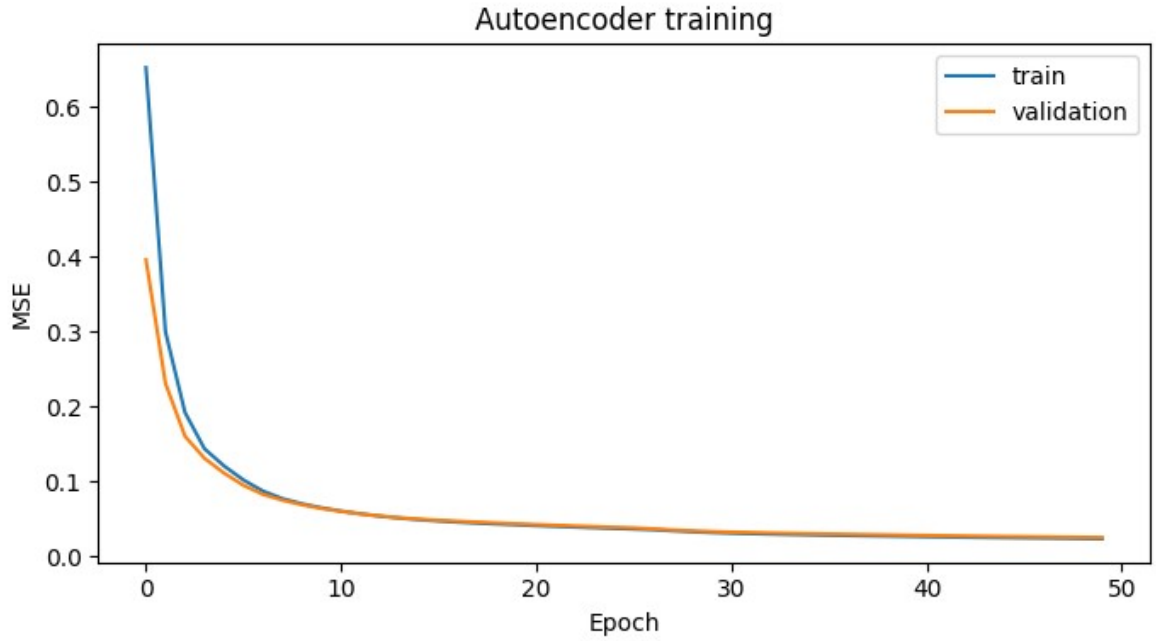


Figure 8 – Dynamics of MSE on train and validation during Autoencoder training

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Balanced XGBoost	82.43%	54.70%	68.78%	54.21%
AE-score + Balanced XGBoost	82.28%	54.48%	68.16%	54.01%

Table 5 – Balanced XGBoost and Autoencoder-score + XGBoost on known classes

At the main threshold, the novelty score Zero-Day Detection Rate was 4.16%, Unknown F1 — 5.64%, Normal FPR — 1.15%. At a softer threshold q=0.90, ZDR increases to 11.98% with Normal FPR of 3.16%. This is a more careful compromise compared to the IF-based approach, but the absolute level of Fuzzers detection remains insufficient.

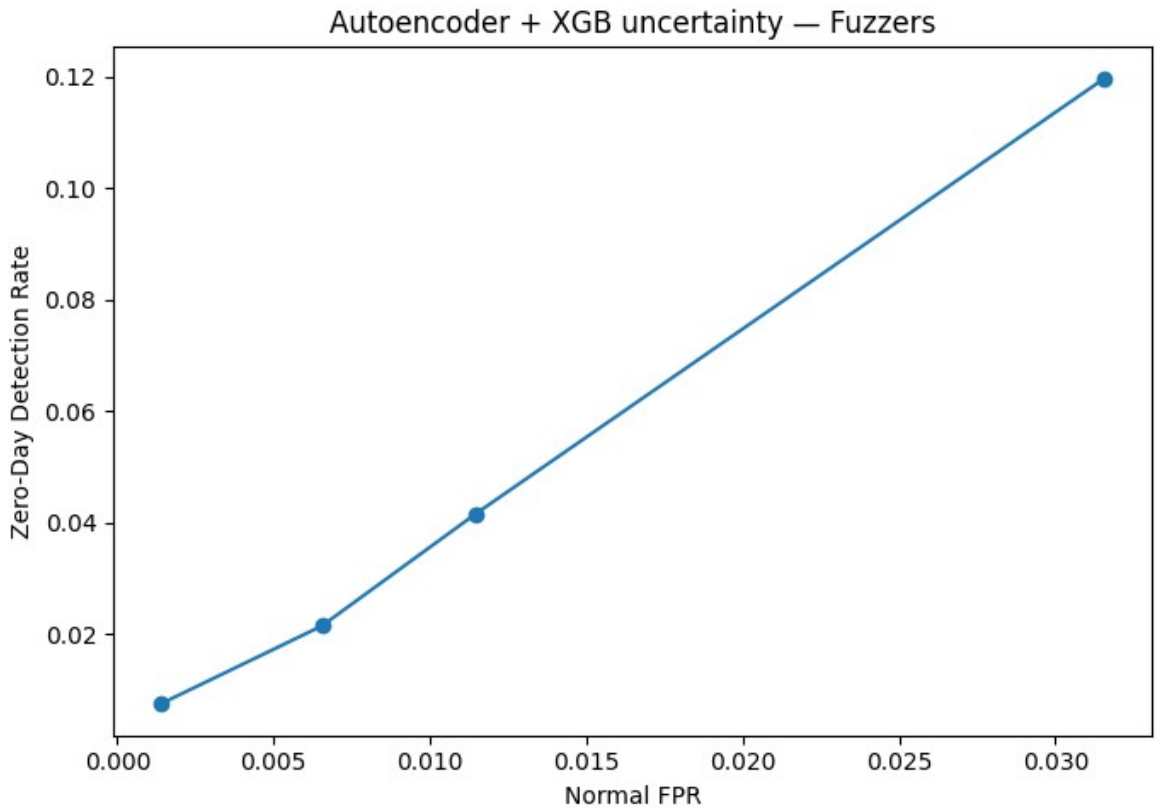


Figure 9 – The trade-off between Normal FPR and Zero-Day Detection Rate for Autoencoder + XGBoost

Threshold q	Normal FPR	ZDR	Unknown Precision	Unknown F1
0.900	3.16%	11.98%	10.50%	11.19%
0.950	1.15%	4.16%	8.77%	5.64%
0.975	0.65%	2.14%	8.43%	3.42%
0.990	0.14%	0.74%	7.55%	1.35%

Table 6 – FPR–ZDR trade-off for Autoencoder + XGBoost

4.5. A summary comparison of three experiments

Model / version	Zero-Day	ZDR	Unknown F1	Normal FPR	Comment
v1 IF + XGBoost	Fuzzers	0.61%	1.15%	0.80%	Fixed thresholds
v2 Calibrated IF + XGBoost	Fuzzers	11.40%	6.84%	5.88%	High share of Known→Unknown
v3 AE + XGBoost (q=0.95)	Fuzzers	4.16%	5.64%	1.15%	Lower FPR
v3 AE + XGBoost (q=0.90)	Fuzzers	11.98%	11.19%	3.16%	A softer threshold

Table 7 – Summary comparison of Zero-Day results for the Fuzzers class

The main conclusion of the current stage is that simply combining a supervised classifier with a general anomaly detector does not solve the Zero-Day recognition problem. Isolation Forest can increase sensitivity, but it quickly raises the number of false positives. Autoencoder provides a more controlled FPR, but Fuzzers remain close to the known feature space and rarely receive the Unknown label. This supports the hypothesis that a specialized Open-Set mechanism is needed — one that takes into account the degree of belonging to known classes, rather than just statistical anomaly. [1, 2, 8, 16, 17]

4.6. The next experiment: Open-Set / EVT

The next stage involves planning an architecture in which XGBoost or an intermediate representation generates estimates of belonging to known classes, and the tails of the distance/confidence distribution are calibrated using the Extreme Value Theory method. The goal of v4 is to achieve a more robust separation of Known and Unknown classes with controlled FPR. After selecting the main Open-Set model, the experiment will be repeated for several held-out attacks (Fuzzers, Exploits, Reconnaissance, DoS, Shellcode). [8, 30, 31]

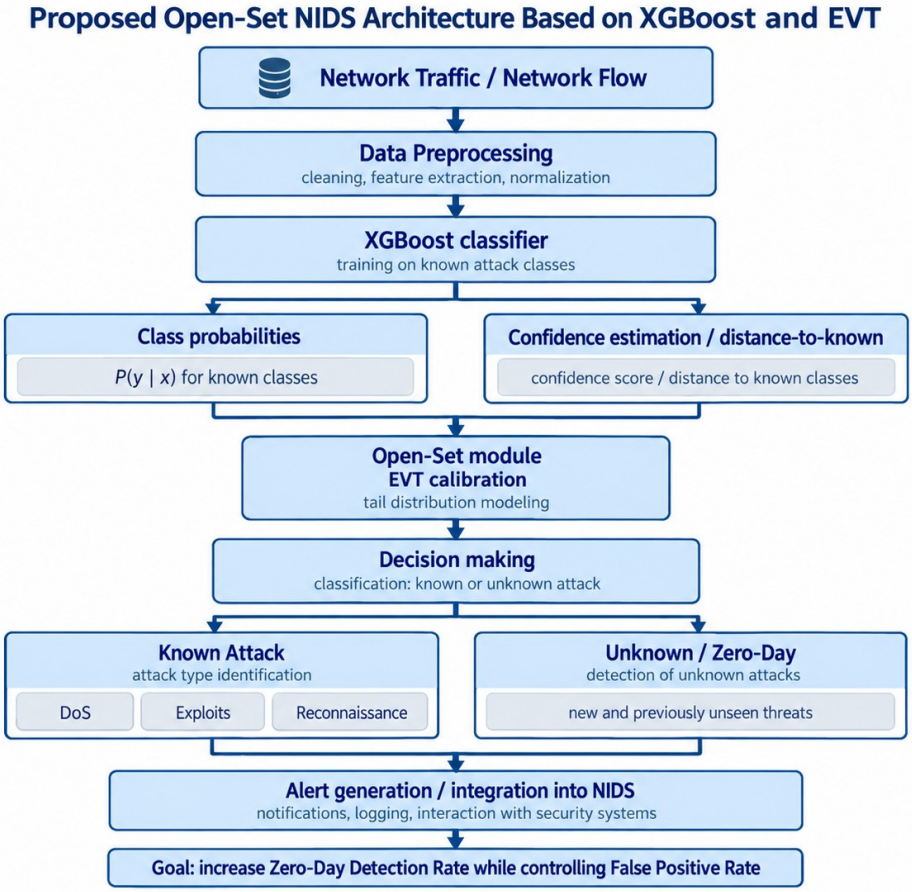


Figure 10 – Proposed Open-Set / EVT architecture

5. Discussion

5.1. The quality of classification of known attacks

On known classes, XGBoost remains the strongest baseline. After balancing, the overall Accuracy decreases compared to the initial run, but Macro Recall increases. This is fundamentally important for NIDS: optimizing only Accuracy may lead to ignoring rare but critical classes.

5.2. Why is Isolation Forest limited?

Isolation Forest is good at detecting statistical outliers, but a Zero-Day attack does not necessarily have to be an extreme outlier relative to normal traffic. Fuzzers partially overlap with the known feature space, so an increase in ZDR is achieved mainly by expanding the anomaly zone and is accompanied by an increase in FPR. [16, 17, 29]

5.3. Why doesn’t the Autoencoder solve the problem completely?

The autoencoder models the normal manifold through reconstruction error. A low reconstruction error for an unknown attack is possible if its numerical features are well reproduced by the normal traffic model. Therefore, reconstruction error alone is insufficient for full-fledged Open-Set Recognition. [20, 21]

5.4. The FPR–ZDR compromise

The increase in sensitivity to unknown attacks should be assessed in conjunction with the False Positive Rate. In a real SOC environment, a high FPR leads to alert fatigue and reduces the practical value of the model. Therefore, the optimal point is not selected based on the maximum ZDR, but on the balance of ZDR, Unknown Precision, FPR, and the quality of known classes. [1, 9, 18, 19]

5.5. Practical integration

A promising pipeline may include NetFlow/Zeek as a source of network indicators, AI-NIDS as a module for Known/Unknown detection, and SIEM/SOAR for event aggregation. An additional layer of MITRE ATT&CK can map known attacks and suspicious chains to TTPs, providing SOC analysts with context rather than just a binary alert. [24–26]

6. Limitations of the study

- The current main experiment was conducted on UNSW-NB15; verification on another modern dataset is required;
- The held-out class of Fuzzers is a model of a Zero-Day scenario, not proof of detecting a real unknown exploit;
- There is currently no verification on real streaming traffic and production infrastructure;
- In the current versions v1–v3, one main held-out class is analyzed; repetition is required for several types of attacks;
- The results depend on preprocessing, the composition of features, and the method of calibrating the Unknown threshold;
- For publication-level evaluation, cross-dataset validation is required, and, preferably, a statistical assessment of variability across several random seeds.. [9–11]

7. Conclusion

This paper provides a comprehensive analysis of modern approaches to modeling and detecting network attacks using artificial intelligence and machine learning algorithms. Supervised, anomaly-based, hybrid, and open-set approaches to building network intrusion detection systems are considered, and an experimental evaluation of Random Forest, XGBoost, Isolation Forest, and Autoencoder is carried out using the UNSW-NB15 dataset. To model previously unknown attacks, the Leave-One-Attack-Class-Out protocol is used, in which the selected attack class is completely excluded from the training dataset and treated as a Zero-Day at the testing stage.

The results obtained show that high classification efficiency for known attacks does not automatically mean a high ability of the model to detect unknown threats. Among the basic supervised models, XGBoost demonstrated the most consistent results on known classes. At the same time, adding an anomaly score obtained using Isolation Forest had virtually no impact on the quality of classification of known attacks. After calibrating Isolation Forest, the sensitivity to the held-out Fuzzers class increased, but this was accompanied by a significant increase in False Positive Rate and the number of known objects erroneously classified into the Unknown category.

Using an Autoencoder made it possible to achieve a more controlled trade-off between the detection rate of Zero-Day attacks and the number of false positives. However, even when combining reconstruction error with the uncertainty of XGBoost, the absolute detection rate for the unknown Fuzzers class remained insufficient. Thus, it has been experimentally confirmed that statistical anomaly and high reconstruction error are not sufficient criteria for classifying a network event as belonging to the unknown class. An unknown attack may be

close to the distribution of normal or previously known malicious traffic and therefore may not be detected as a pronounced outlier by classical anomaly detection methods.

The overall result of the study is that in order to effectively detect Zero-Day attacks, it is advisable to consider the NIDS task not only as an anomaly search task, but also as an Open-Set Recognition task. A promising architecture should simultaneously take into account the probability that a network event belongs to known classes, the degree of confidence of the classifier, and the mechanism for rejecting forced classification if the object does not belong sufficiently to the trained class space.

The conducted experiments formed the basis for the next stage of the research — the development of an Open-Set mechanism using XGBoost and Extreme Value Theory, aimed at increasing the Zero-Day Detection Rate while maintaining a controlled False Positive Rate. To confirm the robustness of the proposed approach, further experiments should include repeating Leave-One-Attack-Class-Out for several types of attacks, cross-dataset validation, and testing the model on streaming network traffic.

Thus, the main conclusion of the study is that the most promising direction is not simply combining a supervised classifier with an anomaly detector, but building a hybrid Open-Set NIDS capable of maintaining high accuracy in recognizing known attacks while simultaneously identifying previously unknown network traffic as a separate category — Unknown.

Список литературы

[1] Guo Y. A review of Machine Learning-based zero-day attack detection: Challenges and future directions. Computer Communications, 198 (2023), 175–185. **DOI:** 10.1016/j.comcom.2022.11.001

[2] Soltani M., Ousat B., Jafari Siavoshani M., Jahangir A.H. An adaptable deep learning-based intrusion detection system to zero-day attacks. Journal of Information Security and Applications, 76 (2023), 103516. **DOI:** 10.1016/j.jisa.2023.103516

[3] Dai Z., Por L.Y., Chen Y.-L., Yang J., Ku C.S., Alizadehsani R., Pławiak P. An intrusion detection model to detect zero-day attacks in unseen data using machine learning. PLOS ONE, 19(9) (2024), e0308469. **DOI:** 10.1371/journal.pone.0308469

[4] Long Z., Yan H., Shen G., Zhang X., He H., Cheng L. A Transformer-based network intrusion detection approach for cloud security. Journal of Cloud Computing, 13 (2024), 5. **DOI:** 10.1186/s13677-023-00574-9

[5] Zhong M., Lin M., Zhang C., Xu Z. A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges. Computers & Security, 141 (2024), 103821. **DOI:** 10.1016/j.cose.2024.103821

[6] Sewak M., Sahay S.K., Rathore H. Deep Reinforcement Learning in the Advanced Cybersecurity Threat Detection and Protection. Information Systems Frontiers, 25(2) (2023), 589–611. **DOI:** 10.1007/s10796-022-10333-x

[7] Sarhan M., Layeghy S., Gallagher M., Portmann M. From zero-shot machine learning to zero-day attack detection. International Journal of Information Security, 22 (2023), 947–959. **DOI:** 10.1007/s10207-023-00676-0

[8] Baye G., Silva P., Broggi A., Fiondella L., Bastian N.D., Kul G. Performance Analysis of Deep-Learning Based Open Set Recognition Algorithms for Network Intrusion Detection Systems. NOMS 2023 – IEEE/IFIP Network Operations and Management Symposium (2023), 1–6. **DOI:** 10.1109/NOMS56928.2023.10154410

[9] Cantone M., Marrocco C., Bria A. Machine Learning in Network Intrusion Detection: A Cross-Dataset Generalization Study. IEEE Access, 12 (2024), 144489–144508. **DOI:** 10.1109/ACCESS.2024.3472907

[10] Mishra S., Alshammari N.S., Hussain H., Alfahidah R.A. A cross-dataset harmonized intrusion detection framework with statistically validated multi-model learning. PLOS ONE, 21(4) (2026), e0346982. **DOI:** 10.1371/journal.pone.0346982

[11] Zhao Q., Zhong S. TriFusion-IDS: A Multimodal Graph-Tabular-Text Contrastive Framework for Cross-Dataset Intrusion Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 40(19) (2026), 16433–16440. **DOI:** 10.1609/aaai.v40i19.38682

[12] Hossain M.M., Turja S.D., Tasnim S., Juboraj M.F.-U.-A., Hossain M.I. A cross-dataset based zero-day intrusion detection system by integrating siamese network and reinforcement learning. ICT Express, 12(3) (2026), 752–757. **DOI:** 10.1016/j.icte.2026.05.001

[13] Xu R., Wu G., Wang W., Gao X., He A., Zhang Z. Applying self-supervised learning to network intrusion detection for network flows with graph neural network. Computer Networks, 248 (2024), 110495. **DOI:** 10.1016/j.comnet.2024.110495

[14] Zhang A., Zhao Y., Zhou C., Zhang T. ResACAG: A graph neural network based intrusion detection. Computers & Electrical Engineering, 122 (2025), 109956. **DOI:** 10.1016/j.compeleceng.2024.109956

[15] Xi C., Wang H., Wang X. A novel multi-scale network intrusion detection model with transformer. Scientific Reports, 14 (2024), 23239. **DOI:** 10.1038/s41598-024-74214-w

[16] Wang W., Jian S., Tan Y., Wu Q., Huang C. Robust unsupervised network intrusion detection with self-supervised masked context reconstruction. Computers & Security, 128 (2023), 103131. **DOI:** 10.1016/j.cose.2023.103131

[17] Bovenzi G., Aceto G., Ciuonzo D., Montieri A., Persico V., Pescapé A. Network anomaly detection methods in IoT environments via deep learning: A fair comparison of performance and robustness. Computers & Security, 128 (2023), 103167. **DOI:** 10.1016/j.cose.2023.103167

[18] dos Santos R.R., Viegas E.K., Santin A.O., Tedeschi P. Federated learning for reliable model updates in network-based intrusion detection. Computers & Security, 133 (2023), 103413. **DOI:** 10.1016/j.cose.2023.103413

[19] Abdel Wahab O. Intrusion Detection in the IoT Under Data and Concept Drifts: Online Deep Learning Approach. IEEE Internet of Things Journal, 9(20) (2022), 19706–19716. **DOI:** 10.1109/JIOT.2022.3167005

[20] Shiomoto K. Network Intrusion Detection System Based on an Adversarial Auto-Encoder with Few Labeled Training Samples. Journal of Network and Systems Management, 31 (2023), 5. **DOI:** 10.1007/s10922-022-09698-w

[21] Lopes I.O., Zou D., Abdulqadder I.H., Ruambo F.A., Yuan B., Jin H. Effective network intrusion detection via representation learning: A Denoising AutoEncoder approach. Computer Communications, 194 (2022), 55–65. **DOI:** 10.1016/j.comcom.2022.07.027

[22] Foley M., Hicks C., Highnam K., Mavroudis V. Autonomous Network Defence Using Reinforcement Learning. Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security (2022), 1252–1254. **DOI:** 10.1145/3488932.3527286

[23] Morris A., Procter R., Wallbank C. Evaluating Reinforcement Learning Agents for Autonomous Cyber Defence. Applied AI Letters, 6(3) (2025), e125. **DOI:** 10.1002/ail2.125

[24] Sun H., Shu H., Kang F., Zhao Y., Huang Y. Malware2ATT&CK: A sophisticated model for mapping malware to ATT&CK techniques. Computers & Security, 140 (2024), 103772. **DOI:** 10.1016/j.cose.2024.103772

[25] Branescu I., Grigorescu O., Dascalu M. Automated Mapping of Common Vulnerabilities and Exposures to MITRE ATT&CK Tactics. Information, 15(4) (2024), 214. **DOI:** 10.3390/info15040214

[26] Ahn G., Jang J., Choi S., Shin D. Research on Improving Cyber Resilience by Integrating the Zero Trust Security Model With the MITRE ATT&CK Matrix. IEEE Access, 12 (2024), 89291–89309. **DOI:** 10.1109/ACCESS.2024.3417182

[27] Moustafa N., Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference (MilCIS) (2015), 1–6. **DOI:** 10.1109/MilCIS.2015.7348942

[28] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), 785–794. **DOI:** 10.1145/2939672.2939785

[29] Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining (ICDM) (2008), 413–422. **DOI:** 10.1109/ICDM.2008.17

[30] Bendale A., Boult T.E. Towards Open Set Deep Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), 1563–1572. **DOI:** 10.1109/CVPR.2016.173

[31] Scheirer W.J., de Rezende Rocha A., Sapkota A., Boult T.E. Toward Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(7) (2013), 1757–1772. **DOI:** 10.1109/TPAMI.2012.256

[32] Vitorino J., Praça I., Maia E. SoK: Realistic adversarial attacks and defenses for intelligent network intrusion detection. Computers & Security, 134 (2023), 103433. **DOI:** 10.1016/j.cose.2023.103433

[33] Haque A., Soliman H. A Transformer-Based Autoencoder with Isolation Forest and XGBoost for Malfunction and Intrusion Detection in Wireless Sensor Networks for Forest Fire Prediction. Future Internet, 17(4) (2025), 164. **DOI:** 10.3390/fi17040164

Appendix A. Outline of diagrams, tables, and graphs for the final article

Below is the complete visual plan. The finished experimental drawings have already been inserted into the main text; the missing diagrams and final graphs will be created after v4 is completed and the series of hold-out experiments are carried out.

№	Material	Type	Status	Place
Figure 1	General research scheme	Scheme	It needs to be done.	After the introduction
Figure 2	Isolation Forest + XGBoost	Architecture diagram	It needs to be done.	Section 3.4
Figure 3	Autoencoder + XGBoost	Architecture diagram	It needs to be done.	Section 3.5
Figure 4	Leave-One-Attack-Class-Out	Protocol diagram	It needs to be done.	Section 3.6
Figure 5	Baseline vs Hybrid	Bar chart	Ready	Section 4.1
Figure 6	Confusion Matrix IF + XGBoost	Heatmap	Ready	Section 4.1
Figure 7	IF: FPR vs ZDR	Line chart	Ready	Section 4.3
Figure 8	Autoencoder train/validation MSE	Training curve	Ready	Section 4.4
Figure 9	AE + XGB: FPR vs ZDR	Line chart	Ready	Section 4.4
Figure 10	Open-Set / EVT architecture	Architecture diagram	After v4	Section 4.6
Figure 11	Zero-Day Detection Rate by Attack Type	Bar chart	After the hold-out series	Section 4.7
Figure 12	Known F1 vs Zero-Day ZDR vs FPR	Summary schedule	After v4	Results/Discussion
Figure 13	Cross-dataset performance	Bar/heatmap	Later	After the second dataset

Table A1 – Visual Materials Plan

№	Title	Status / action
Table 1	Review of relevant studies	Done: accurate bibliographic records have been added, including authors, publication year, and DOI.
Table 2	UNSW-NB15 classes	Ready in terms of structure; add counts for train/test.
Table 3	Baseline results v1	Ready
Table 4	IF FPR–ZDR trade-off	Ready
Table 5	Balanced XGBoost vs AE-hybrid	Ready
Table 6	AE FPR–ZDR trade-off	Ready
Table 7	Summary comparison of Zero-Day v1–v3	Ready
Table 8	Open-Set v4 vs baselines	After v4
Table 9	Hold-out results for 5 attacks	After a series of experiments
Table 10	Comparison with existing works	After the final model selection

Table A2 – Table of Contents of the Article

Appendix B. What needs to be done next

- create four clean scientific workflows: general workflow, IF+XGBoost, AE+XGBoost, and Leave-One-Attack-Class-Out;
- run v4 Open-Set/EVT and record Known F1, ZDR, Unknown F1, and Normal FPR;
- repeat the held-out experiment for Exploits, Reconnaissance, DoS, and Shellcode;
- build the final bar chart of Zero-Day Detection Rate by Attack Type;
- conduct at least a few repetitions/seeds or justify a deterministic protocol.
- After selecting the main model, add a second dataset (for example, CICIDS2017) to test generalization;
- Check the final compliance of the Zotero bibliography and the cited references before submitting the article;
- After completing the results, rewrite the abstract and conclusion with the final numerical values.

DETECTING AND MITIGATING LOAD TELEMETRY FALSIFICATION IN GRU-BASED PREDICTIVE RESOURCE ALLOCATION

Kabilov R.I.
e-mail: raxmetullakabilov@gmail.com
Data Communication Networks and Systems Department Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
Reypnazarov E.N.
e-mail: e.reypnazarov@tuit.uz
Data Communication Networks and Systems Department Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
Saymanov I.M.
e-mail: saymanov_i@nuu.uz
Department of Information Security, National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan

Abstract. *Predictive resource allocators built around a GRU model can be misled when the load reports feeding them are falsified. We study this problem in a simulated three-group queueing system, where a queue-balance detector cross-checks each reported arrival against the backlog change and trusted service count for that interval; reports that fail this check are corrected before they reach later predictions. Under a sustained +3 service-unit falsification, simply averaging the signed residuals over five intervals pushed detection from 19.2% up to 99.8%, while the false-positive rate stayed close to target — 1.043% and 1.210% for the two settings. Across ten independent realizations, this correction brought the mean attack-induced backlog down from 1.746 to 0.018 service units. The benefit was not unconditional, though: stronger measurement noise and misaligned timing both degraded detection in our sensitivity tests. Taken together, the results support telemetry verification, but only under the assumptions tested here — synthetic traffic, trusted service counters, and properly aligned measurements.*

Keywords: *predictive resource allocation; gated recurrent unit; telemetry falsification; queue-balance detection; attack mitigation.*

Introduction

Radio access networks increasingly rely on AI to guide resource allocation — power control, physical resource block assignment, network slicing, and similar tasks all draw on it [1]. What makes these systems work is the stream of measurements reported by network elements, so if that stream can't be trusted, neither can the decisions built on top of it. This isn't a hypothetical concern: falsified key performance indicators have already been shown to steer traffic in O-RAN deployments [2]. A predictive allocator is particularly exposed here, since altered load reports distort its forecasts and push resources toward groups that don't actually need them — and the groups being shortchanged see their queues grow as a result.

One approach to detecting malicious telemetry is to learn the normal behavior of network measurements. For example, MARRS uses LSTM autoencoders and information from neighboring cells to identify suspicious reports in an O-RAN traffic-steering setting [2]. Here, we investigate a different basis for verification: the balance between arriving workload, changes in queue occupancy, and completed service. In other words, what a node reports as its arrivals should line up with what the queue and service measurements say actually happened. This check requires trusted service counters and measurements referring to the same interval. Measurement noise complicates the comparison, particularly when the falsification is small.

We evaluate this approach in a simplified system where three traffic groups share a fixed service capacity. A gated recurrent unit (GRU) network [3] predicts their workloads, and an allocator distributes resources using these predictions and the current backlogs. We compare bounded gradient-based perturbations with constant positive falsification, then test

instantaneous and window-averaged queue-balance detectors. Suspicious reports are corrected before they enter subsequent prediction windows. We track two things throughout: how well the attack is detected, and how much extra backlog it causes. Beyond the main attack scenario, we also stress-test the approach against traffic bursts, measurement noise, and delayed observations, which is where verification starts to help less. All of this is tied to the specific simulation and measurement setup we use — it says something about when the method works, not about how it would fare in a real deployed network.

System and Threat Models

System Model

Our setup has three traffic groups sharing one resource pool of capacity C, with time split into discrete intervals and a single allocation decision made each interval. For group i we write $A_i(t)$ for the workload arriving during interval t and $Q_i(t)$ for its backlog at the start of that interval. Everything here is expressed in abstract service units rather than tied to physical resource blocks or a specific radio transmission rate.

At the start of each interval, the controller hands each group a nonnegative allocation $a_i(t)$, with the total across groups capped at C. Since arrivals can be served the moment they come in, how much actually gets served depends on both the workload on hand and the allocation assigned to it. This gives us the completed service and the backlog rolled into the next interval, shown in (1).

$$\begin{aligned} S_i(t) &= \min\{Q_i(t) + A_i(t), a_i(t)\} \\ Q_i(t+1) &= Q_i(t) + A_i(t) - S_i(t) \end{aligned} \tag{1}$$

We start every queue empty, and buffer capacity is treated as unlimited. Packet losses are not modeled, and unused allocations are not redistributed within the same interval. The allocator uses the current local backlog, while the detector receives separate queue measurements that may contain noise. These assumptions allow us to examine the effect of falsified load reports without introducing radio-channel variability.

Prediction-Based Resource Allocation

To forecast the next interval's workload for all three groups, we feed the GRU the preceding 20 reports. The vector x(t) holds whatever reports survive verification and correction (when these are enabled), and it's from this history that the model produces the workload estimates in (2).

$$\widehat{A}(t) = f_\theta\left(x(t-20), \ldots, x(t-1)\right) \tag{2}$$

Here f is the GRU predictor itself, with θ its trained parameters, and inputs are standardized using statistics drawn only from the training data. Predictions are then converted back to service units, with negative values set to zero.

For each group, the allocator adds the predicted arrivals to the current backlog to obtain its allocation demand. When the combined demand is within capacity, each group receives its requested allocation. Otherwise, the allocations are scaled proportionally to satisfy the capacity constraint, as shown in (3).

$$\begin{aligned} D_i(t) &= Q_i(t) + \max\left\{0, \widehat{A}_i(t)\right\} \\ a_i(t) &= D_i(t) \min\left\{1, \frac{C}{\sum_{j=1}^{3} D_j(t)}\right\} \end{aligned} \tag{3}$$

If the combined allocation demand is zero, all allocations are set to zero. The GRU is trained on benign traffic, and its weights remain fixed throughout the attack and mitigation experiments. As a baseline, we use a persistence predictor, which takes the most recent reported workload as the estimate for the next interval.

Threat Model

The attacker can alter the arrival reports of the first traffic group without changing its actual workload. The other groups' reports, the GRU weights, the allocator's local queue state, and the queue and service measurements used for verification remain outside the attacker's

control. Each falsified report must be nonnegative and differ from the actual arrival value by no more than ε service units, as specified in (4).

$$\widetilde{A}_1(t) = A_1(t) + \delta(t), \quad |\delta(t)| \le \varepsilon, \quad \widetilde{A}_1(t) \ge 0 \tag{4}$$

The attacker modifies each selected report before it enters the history used for the next prediction. Unless corrected, that report remains in the history and can influence subsequent forecasts. The objective is to increase the first group's predicted demand, leading to excess allocation to that group and potentially increasing the backlogs of the other two groups.

We compare two strategies. The constant-bias strategy adds ε to each selected report. The white-box, FGSM-style strategy [4] computes the gradient of the first group's forecast with respect to its latest report and applies a perturbation of magnitude ε in the gradient's sign direction. The modified value is clipped at zero. Gradients are calculated using the previously modified history, without access to future arrivals.

For the mitigation experiments, we generate the attack reports in advance and replay the same sequence in protected and unprotected runs. This allows us to compare their effects under different protection settings. The attacker does not adapt its reports to the detector or the correction mechanism.

Detection and Mitigation Method

Queue-Balance Residual

At its core, the detector simply checks reported arrivals against the workload we can reconstruct from how the queue evolved and how much was served. At the end of each interval, it receives queue measurements for the beginning and end of that interval, together with a trusted service count. All measurements must refer to the same interval. Using Q^{obs} to denote the measured backlog, we define the signed residual for each group in (5).

$$e_i(t) = Q_i^{(obs)}(t+1) - Q_i^{(obs)}(t) + S_i(t) - \widetilde{A}_i(t) \tag{5}$$

When reports are truthful and measurements are exact, queue conservation forces this residual to zero. Queue-measurement errors introduce fluctuations around zero, while an inflated arrival report contributes a negative residual.

We evaluate two detection rules. The instantaneous rule uses the current residual, whereas the windowed rule first averages the signed residuals over the latest W intervals. Both rules then take the absolute value and use the largest value across the three groups as the interval-level score, as shown in (6).

$$z_W(t) = \max_{i=1,2,3} \left| \left(\frac{1}{W}\right) \sum_{k=0}^{W-1} e_i(t-k) \right| \tag{6}$$

We flag an interval whenever its score crosses a threshold, and that threshold comes from separate, benign calibration data rather than the test set itself. Given a target false-positive rate α, this threshold is just the corresponding empirical quantile — equation (7) spells out both the threshold and the resulting alarm rule.

$$\begin{gathered} \tau_W = Quantile_{(1-\alpha)} Z_W^{cal} \\ z_W(t) > \tau_W \text{ indicates an alarm.} \end{gathered} \tag{7}$$

We set α to 0.01 and calibrate separate thresholds for W = 1 and W = 5. No attack samples are used to select these thresholds. Windowed detection starts only after W residuals are available. Averaging signed residuals can reduce measurement fluctuations while preserving a sustained bias, but the benefit depends on the noise structure and temporal alignment. The target false-positive rate applies to the calibration procedure; the observed rate on test data may differ.

Report Correction

Each traffic group is corrected on its own. Whenever a group's window-averaged residual crosses the calibrated threshold in absolute value, we swap out its current report for an arrival estimate rebuilt from the queue measurements and service count. Negative reconstructed values are set to zero. Reports that do not trigger the detector are retained, as expressed in (8).

$$x_i(t) = \begin{cases} \max\left\{0, \quad Q_i^{(obs)}(t+1) - Q_i^{(obs)}(t) + S_i(t)\right\}, & \left|\left(\frac{1}{W}\right)\sum_{k=0}^{W-1} e_i(t-k)\right| > \tau \\ \widetilde{A}_i(t), & otherwise \end{cases} \tag{8}$$

The retained or corrected report enters the history used for subsequent GRU predictions. Because verification only happens once the interval has already closed, correction can never touch the allocation that was already made — it's too late for that. Earlier entries in the history and the trained model's weights are likewise left untouched.

What does carry the original signal forward is the detection window itself: it keeps residuals computed from the reports as they originally arrived, corrected or not. This prevents the correction step from removing evidence of falsification from subsequent checks. Although detection averages residuals over several intervals, the replacement value is reconstructed from the current interval only and may therefore contain measurement noise. We assess correction under both attacked and benign traffic to measure its effect on backlog, including any harm caused by false alarms.

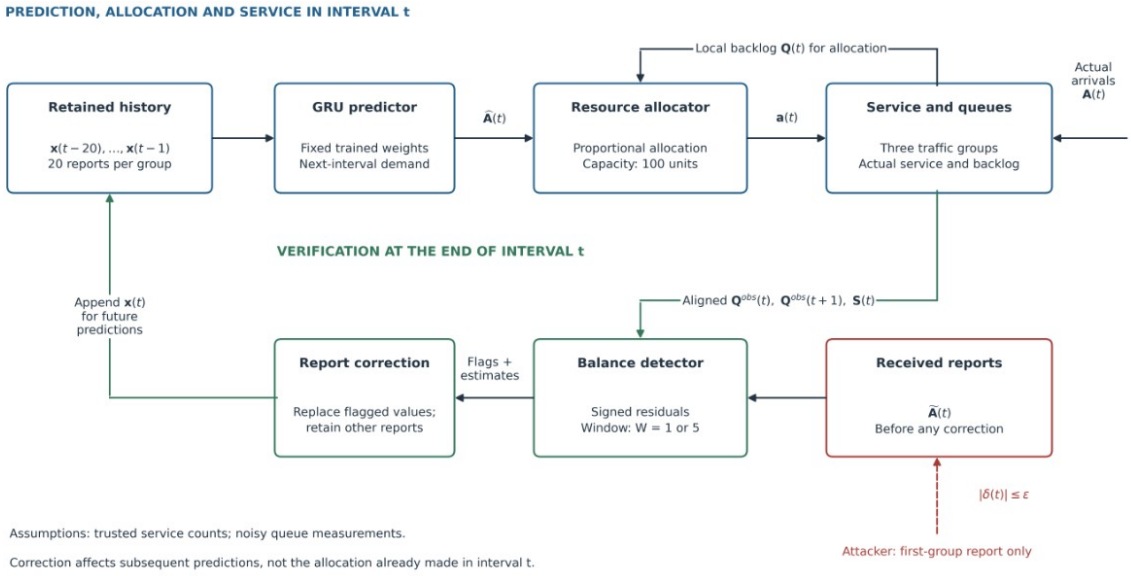


Figure 1. GRU-based predictive resource allocation with queue-balance telemetry verification and report correction

Figure 1 shows the timing of allocation and verification. Each report gets checked only after that interval's service has already happened, so any correction feeds forward into future predictions and allocations rather than the current one.

Experimental Setup and Results

Experimental Setup

We built the experiments in Python with NumPy and TensorFlow/Keras. Each group's synthetic traffic came from a constant mean plus a sinusoidal component plus independent Gaussian noise — mean loads of 18, 25, and 33 service units, sinusoidal amplitudes of 4, 5, and 6, and periods of $2\pi \times 43$, $2\pi \times 67$, and $2\pi \times 91$ intervals, respectively, with the Gaussian noise set to a standard deviation of 2 units and negative arrivals clipped to zero.

For the GRU experiments specifically, we scaled arrivals by 1.25 so the nominal aggregate load sat at 95% of capacity, then generated a 12,000-interval sequence and split it chronologically into 7,200 training, 2,400 validation, and 2,400 test intervals. Since each input window only looks at reports before its prediction target, and we used a 20-interval history, this left 7,180 usable samples for training. Standardization used statistics from the training period only.

We trained the GRU with mean squared error and the Adam optimizer [5]. Early stopping monitored validation loss and restored the weights from the best monitored epoch. The model weights were then fixed for all attack and mitigation tests. Table 1 lists the main experimental parameters.

Table 1. Main experimental parameters

Parameter	Value
Number of traffic groups	3
Service capacity	100 service units per interval
Nominal load in GRU experiments	95%
GRU input window / forecast horizon	20 intervals / 1 interval
Predictor architecture	GRU with 32 units; linear output layer with 3 units
Trainable parameters	3,651
Optimizer / learning rate	Adam / 0.001
Batch size / maximum epochs	64 / 60
Early stopping	Patience: 8 epochs; minimum improvement: 0.0001
Test duration	2,400 intervals
Main attack	First-group reports; bound of 3 service units; 500 consecutive reports
Detector windows	1 and 5 intervals
Nominal queue-measurement noise	Independent Gaussian errors; standard deviation: 1 service unit
Calibration target FPR	1%
Mitigation repetitions	10 traffic and measurement-noise realizations
Noise and timing sensitivity tests	30 measurement-noise realizations per condition

We measured forecasting accuracy using mean absolute error and allocation performance using queue backlog. To assess the attack's impact on competing traffic, we calculated the mean combined backlog of groups 2 and 3 over the 500-interval attack decision period. Attack-induced backlog was the difference between attacked and benign runs using the same protection setting. Detection was evaluated using the true-positive rate (TPR), the fraction of attacked intervals flagged, and the false-positive rate (FPR), the fraction of benign intervals flagged.

Across the repeated mitigation tests, we varied the traffic phases, the random fluctuations in arrivals, and the queue-measurement noise, but kept the same trained GRU throughout. 95% confidence intervals were then computed from the replicate-level measurements using the Student t distribution.

Forecasting and Attack Impact

On the benign test sequence, the GRU achieved a mean absolute error of 2.071 service units per interval, compared with 2.806 for the persistence predictor—a reduction of approximately 26.2%. The improvement in allocation performance was smaller: mean total backlog fell from 205.999 to 203.208 service units, or 1.35%. When the same allocator was supplied with the actual arrivals, mean backlog was still 197.818. These results indicate that prediction error accounted for only part of the queueing in this scenario; periods when demand exceeded capacity also contributed substantially.

During the 500-interval attack decision period, perturbations bounded by 3 service units increased the first group's mean forecast from 22.069 to 24.586 and its mean allocation from 22.674 to 23.831. The mean combined backlog of groups 2 and 3 rose from 15.500 to 17.362 service units, an increase of approximately 12.0%. Across the full 2,400-interval test, the increase in mean total backlog was smaller, from 203.208 to 203.650, or approximately 0.22%. The attack's effect was therefore more apparent when examining the competing groups during the attack period than when averaging across all groups and the entire test.

The gradient-based attack selected a positive perturbation for all 500 modified reports. It therefore produced the same report sequence as constant +3 falsification, giving identical forecasts and allocation outcomes. In this experiment, gradient information provided no additional advantage over a constant positive bias. This finding is specific to the tested model, attack objective, and perturbation constraints.

Detection and Mitigation Performance

When queue-measurement noise had a standard deviation of 1 service unit, the five-interval detector picked up sustained +3 falsification far more reliably than the instantaneous one. With thresholds calibrated on 100,000 benign intervals for a 1% target FPR, the test TPRs came out to 19.2% for the instantaneous detector and 99.8% for the five-interval one, while their observed FPRs stayed at 1.043% and 1.210%. So window averaging bought real sensitivity without pushing false positives away from the calibration target.

The mitigation tests used the initial benign-calibrated thresholds: approximately 4.101 for instantaneous detection and 0.803 for five-interval detection. In the single test realization, the attack increased the mean combined backlog of groups 2 and 3 from 15.500 to 17.362 service units. Instantaneous correction reduced it to 16.783, while five-interval correction brought it to 15.506, close to the benign reference. In the corresponding runs without an attack, backlogs were 15.503 and 15.500, respectively.

To assess whether the mitigation benefit persisted across different realizations, we repeated the five-interval experiment with ten additional traffic and measurement-noise realizations under constant +3 falsification. The trained GRU and the initial windowed threshold remained fixed. Table 2 summarizes these results.

Table 2. Mitigation performance under sustained +3 service-unit falsification (five-interval detector, 10 realizations)

Metric	**Mean ± 95% CI half-width**
Attack-induced backlog without mitigation, service units	1.746 ± 0.727
Attack-induced backlog with mitigation, service units	0.018 ± 0.021
Backlog reduction due to mitigation under attack, service units	1.718 ± 0.715
Backlog change due to mitigation under benign traffic, service units	0.010 ± 0.023
TPR, %	99.820 ± 0.081
Benign FPR, %	1.411 ± 0.177

The backlog metrics in Table 2 refer to groups 2 and 3 over the 500-interval attack decision period, or the matching period in benign runs. For each protection setting, attack-induced backlog is calculated relative to its own benign baseline. Mean attack-induced backlog fell from 1.746 service units without protection to 0.018 with five-interval correction. Comparing these means gives a reduction of approximately 99% in the additional backlog caused by the attack, not in total system backlog.

The confidence interval for the backlog change under benign traffic included zero. So across these repetitions, correction didn't reliably drive up benign backlog either way. Taken as a whole, this supports using report correction against the kind of sustained falsification we tested — provided service counters are trusted and measurements stay aligned. What it doesn't show is protection against an attacker who can tamper with those counters, or who adapts their reports once they know the defense is there.

Sensitivity Analysis and Limitations

We also wanted to see how the detector handles a legitimate traffic change, so we bumped the first group's actual arrivals by 20 service units for 200 intervals while keeping its reports honest throughout. The five-interval detector's FPR came out to 1.500% during that burst and 1.628% across the full test, and mean total backlog barely moved — 1062.639 service units without protection versus 1062.634 with correction. So correction essentially did nothing to backlog here, which makes sense given queue conservation: when reports stay accurate, a real increase in arrivals just doesn't create the kind of balance mismatch the detector looks for.

Measurement noise had a stronger effect on detection. When thresholds calibrated at a noise standard deviation of 1 were used at a standard deviation of 5, the five-interval detector’s FPR rose to 91.20%. Calibrating separate thresholds for each noise level kept observed FPRs near 1%, but sensitivity still declined. The five-interval detector’s TPR fell from 99.82% at a

standard deviation of 1 to 72.01% at 3 and 21.27% at 5. These tests assumed that the noise level was known; automatic online threshold adaptation was not evaluated.

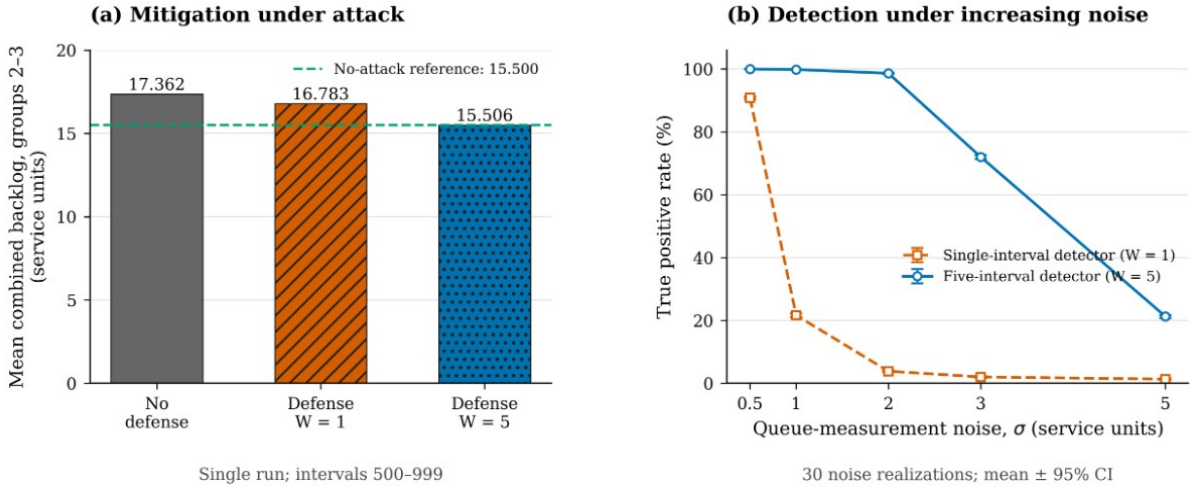


Figure 2. Mitigation performance and detection sensitivity under first-group telemetry falsification (+3 service units): (a) mean combined backlog of groups 2–3 over intervals 500–999 in a single run; (b) mean TPR with 95% confidence intervals across 30 noise realizations, with thresholds calibrated separately at each noise level to a target FPR of 1%

Looking at Figure 2(a), five-interval correction pulled the competing groups' mean backlog back close to the benign reference, while instantaneous correction still left a noticeably bigger attack footprint — consistently catching the sustained falsification seems to be what limits its reach into later forecasts. Figure 2(b), though, shows this isn't free: the benefit hinges on measurement quality. Recalibrating kept false alarms in check, but as noise got stronger, small falsifications simply became harder to tell apart from ordinary benign fluctuations.

We also tested the effect of delayed queue measurements. Combining these measurements with current service counts and reports raised the five-interval detector’s FPR to 59.83%, 69.09%, and 80.59% for delays of one, two, and five intervals, respectively. Matching all quantities to the same source interval maintained a TPR of approximately 99.83% and an FPR of 0.94%, although the decision then arrived later. We did not evaluate how this delay would affect mitigation or backlog.

The results depend on the assumptions of a fluid queueing model, synthetic traffic, one trained GRU, and trusted service counters. The assumed noise structure also favors window averaging: consecutive queue-measurement errors cancel in the residual sum, leaving only the errors at the window boundaries. This explains part of the observed detection benefit. Radio-channel variability, packet losses, service-counter errors, and attacks that adapt to the defense were not modeled. The experiments therefore identify the conditions under which the method works in this simulation, rather than demonstrate robustness in a deployed network. Testing attacks that explicitly account for the defense remains an important step in assessing adversarial robustness [6].

Conclusion

We examined how falsified load reports affect a GRU-based predictive allocator serving three traffic groups. Although the GRU cut forecasting error by 26.2% relative to a persistence baseline, mean total backlog dropped by just 1.35%, showing that better predictions did not translate into a proportionally better allocation under the tested load. Falsifying the first group's reports increased backlog for competing groups, while the gradient-based attack proved no more effective than constant positive falsification.

Queue-balance verification and report correction substantially reduced this impact: across ten additional realizations, five-interval correction lowered mean attack-induced backlog from

1.746 to 0.018 service units, with a benign FPR of 1.411%. This benefit depends on trusted service counts, aligned measurements, and proper threshold calibration — stronger noise reduced detection sensitivity, and temporal misalignment caused frequent false alarms. Future work should address service-counter errors, delayed correction, more varied traffic, and attackers that adapt to the defense.

References

[1] S. Nouri, M. Karbalaee Motalleb, V. Shah-Mansouri, and S. P. Shariatpanahi, “Generative AI for O-RAN Slicing: A Semi-Supervised Approach with VAE and Contrastive Learning,” arXiv:2401.08861, version 3, 2025. https://doi.org/10.48550/arXiv.2401.08861

[2] E. Aizikovich, D. Mimran, E. Grolman, Y. Elovici, and A. Shabtai, “Rogue Cell: Adversarial Attack and Defense in Untrusted O-RAN Setup Exploiting the Traffic Steering xApp,” arXiv:2505.01816, 2025. https://doi.org/10.48550/arXiv.2505.01816

[3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” Proceedings of EMNLP, pp. 1724–1734, 2014. https://doi.org/10.3115/v1/D14-1179

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” International Conference on Learning Representations (ICLR), 2015. https://arxiv.org/abs/1412.6572

[5] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” International Conference on Learning Representations (ICLR), 2015. https://arxiv.org/abs/1412.6980

[6] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, “On Adaptive Attacks to Adversarial Example Defenses,” arXiv:2002.08347, 2020. https://doi.org/10.48550/arXiv.2002.08347

ҚҰПИЯНЫ БӨЛІСУ СХЕМАЛАРЫ: ӘДІСТЕРІ МЕН ҚОЛДАНЫЛУЫ

Дюсенбаев Д.С., Хомпыш А., Ищукова Е.А.
e-mail: dimash_dds@mail.ru , ardabek@mail.ru , jekky82@mail.ru
ҚР ҒЖБМ «Ақпараттық және есептеуіш технологиялар институты»
Ақпараттық қауіпсіздік зертханасы

Аңдатпа. *Бұл мақалада құпияны бөлісу схемаларына қолданылатын қалдықтар туралы қытай теоремасы, Лагранж интерполяциясының көпмүшелік схемасы, векторлық құпияны бөлісу схемасы және теңдеулер жүйесін шешуге негізделген схемалары талқыланып, солардың негізінде жаңа модификацияланған құпияларды бөлісу схемасы құрастырылды. Құрастырылған модификация схемасы жария түрде тексерілетін құпия бөлісуге жатады. Шамирдің құпияларды бөлісу схемасынан ерекшелігі кез келген k құпия ұстаушы құпияны аша бермейді, қандайда реттеп терілген k құпия ұстаушы құпияны қалпына келтіре алады.*

Кілт сөздер: *Шамирдың құпия бөлісуі (SSS), тексерілетін құпия бөлісуі (VSS), жария түрде тексерілетін құпия бөлісуі (PVSS), қалдықтар туралы қытай теоремасы, Лагранж интерполяциясының көпмүшелік схемасы.*

1 Кіріспе

Қалдықтар туралы қытай теоремасын негізделген Миньот схемасы және Асмут-Блум схемасы, Лагранж интерполяциясының көпмүшелік схемасына негізделген Шамир схемасы [1], векторлық құпияны бөлісу схемасына негізделген Блэкли схемасы, теңдеулер жүйесін шешуге негізделген Карнин-Грин-Хелман схемасы қарастырылды және өзара талқылау жүргізілді. Шамир, Блэкли және Карнин-Грин-Хелман схемалары әртүрлі жүйеде құрылғанмен, құпияны қалпына келтіруде, барлығын бір жүйеге сызықтық теңдеулер жүйесіне әкелуге болады. Осы айтылған үш схемада k айнымалы n сызықтық теңдеулер жүйесін шешуге негізделген (сүйенеді), өйткені n сызықтық теңдеулердің кезкелген k теңдеулері бойынша айнымалыларды анықтай аламыз. Ал егер теңдеулердің саны k-дан кіші болса, шешім көп болып, анықтау қиындайды.

Қалдықтар туралы қытай теоремасын пайдаланғанда, аргументтің модулдерге қатысты немесе керісінше аргументке қатысты модулдерге шектеу қоюға міндеттейді. Миньот және Асмут-Блум схемалары да осындай шектеулерге сүйенеді. Құпияны анықтауда модульдардың саны емес, олардың көбейтіндісі маңызды, сондықтан құпияны қалпына келтіруге n модулдің кез-келген қандайда көбейтінділері құпиядан артық болуы жеткілікті.

Модификациясы ретінде екі схеманың композициялары да қарастырылған. Мысалға, таңдалған құпияны қалдықтар туралы қытай теоремасы бойынша алынған қалдықтарды көпмүшеліктің коэффициенттері ретінде көреміз. Алынған көпмүшелікті Лагранж интерполяциясының көпмүшелік схемасы ретінде қарастырамыз.

Тексерілетін құпия бөлісу (VSS) схемаларын Педерсен [3] және Фельдман [4] ұсынды, дилер тарату хаттамасы кезінде кейбір немесе барлық қатысушыларға жарамсыз акцияларды жібере алмайды және қатысушылар қалпына келтіру хаттамасы кезінде жарамсыз акцияларды жібере алмайды.

Тексерілетін құпия бөлісу (VSS) схемалары үшін кейбір жағдайларда (мысалы, электрондық дауыс беру) екі кемшілік бар: қатысушылар тек өз үлестерін ғана тексере алады, бірақ қатысушылардың дұрыс үлесті алғанын ешкім тексере алмайды. қатысушылар қайтару хаттамасында өз үлестерін босатады, содан кейін босатылған үлестерді тарату хаттамасының нәтижелері бойынша кез келген адам тексере алады. Осы мәселелерді шешу үшін Штадлер (Stadler [6]) және Шенмакерс (Schoenmakers [4]) жария түрде тексерілетін құпия бөлісу (PVSS) схемаларын ұсынды.

Құпияларды бөлісу схемаларында адал емес қатысушылардан (шабуылшылардан) қорғану үшін аутентификациялау тәсілдерін қолдануға мәжбүр етеді [2].

Аутентификация қолданылатын тексерілетін құпия бөлісу (VSS) және жария түрде тексерілетін құпия бөлісу (PVSS) схемалары дискретті логарифм мен Диффи-Хеллман есептерінің қиындықтарына сүйенеді.

2 Модификацияланған құпияны бөлісу схемасы

1) Инициализациялау

Дилер q жай сан және оның g приметивті элементті генерациялау арқылы алады. Дилер (g, q) жұбын барлық барлық құпия ұстаушыларға таратады. Дилер $s \in Z_q$ құпия кілтін қалыптастырады және $x = g^s(modq)$ ашық кілтін есептеп, оны барлық құпия ұстаушыларға жібереді.

Әрбір L_i құпия ұстаушы өзінің $s_i \in Z_q$ құпия кілттерін қалыптастырады және $x_i = g^{s_i}(modq)$ ашық кілтін есептеп, оны дилерге тіргейді.

2) Құпияны бөлісу

Q құпияны ашуға мүмкіндігі бар k санына қатысты таңдалған $p_1, p_2, \cdots, p_k$ жай сандарға қатысты көпмүшеліктің коэффициенттері төмендегідей анықталады:

$$\begin{cases} Q = a_{k-1}(modp_1) \\ Q = a_{k-2}(modp_2) \\ \quad\quad\quad ... \\ Q = a_0(modp_k) \end{cases} \quad (1)$$

$$f(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + ... + a_2x^2 + a_1x + a_0 \quad (2)$$

Әрбір құпия ұстаушыға байланысты t_i мәндері кездейсоқ таңдалып, келесі мәндер $f(t_i) = b_i(mod\ P), i = \overline{1,n}$есептеліп, келесідей $(t_i, b_i, p_j), i = \overline{1,n},\ j = \overline{1,k}$ үштік мәндерді барлық құпия ұстаушыларға таратылады. Мұндағы P жай сандардың көбейтіндісі $P = \prod_{j=1}^{k} p_j$.

Дилер $H_i = H(x_i, t_i, b_i, p_j, x)$ хеш мәнін есептейді және оны сәйкесінше L_i құпия ұстаушыға H_i жібереді.

Дилер әрбір L_i құпия ұстаушының (x_i, H_i) жұбын тіркеп құпияда сақтайды.

3) Құпияны қалпына келтіру

Кез келген k құпия ұстаушы бірігіп құпияны аша бермейді, тек $p_j, j = \overline{1,k}$ барлығы қатысқанда ғана қалпына келтіре алады. Ол үшін k құпия ұстаушы $(t_{i_j}, b_{i_j}, p_j)\ 1 \le i_j \le n, j = \overline{1,k},$ (3) өздерінің құпия үлестері бойынша екі жолмен құпияны қалпына келтіруге болады:

1) Лагранж интерполяциялық көмүшелігін анықтау үшін k құпия ұстаушылардың (3) жиынындағы үштіктің (t_{i_j}, b_{i_j}, p_j) үштіктері бойынша құрылған $f(x) = \sum_{j=1}^{k} b_{i_j} \prod_{j=1, i_l \neq i_j}^{k} \frac{x - t_{i_j}}{t_{i_l} - t_{i_j}} (mod\ P), 1 \le i_j \le n$ (4) формуласын пайдаланып (2) көпмүшеліктің коэффициенттерін анықтаймыз. Анықталған коэффициенттер (1) теңдеулер жүйесін береді және қалтықтар туралы қытай теоремасының көмегімен құпияға қол жеткіземіз.

2) сызықтық теңдеулер жүйесін құру арқылы.

$$\begin{cases} x_{k-1}t_{i_1}^{k-1} + x_{k-2}t_{i_1}^{k-2} + ... + x_2t_{i_1}^2 + x_1t_{i_1} + x_0 = b_{i_1}(modP) \\ x_{k-1}t_{i_2}^{k-1} + x_{k-2}t_{i_2}^{k-2} + ... + x_2t_{i_2}^2 + x_1t_{i_2} + x_0 = b_{i_2}(modP) \\ \dots \\ x_{k-1}t_{i_{k-1}}^{k-1} + x_{k-2}t_{i_{k-1}}^{k-2} + ... + x_2t_{i_{k-1}}^2 + x_1t_{i_{k-1}} + x_0 = b_{i_{k-1}}(modP) \\ x_{k-1}t_{i_k}^{k-1} + x_{k-2}t_{i_k}^{k-2} + ... + x_2t_{i_k}^2 + x_1t_{i_k} + x_0 = b_{i_k}(modP) \end{cases} \quad (5)$$

(5) теңдеулер жүйесі k айнымалы k сызықтық теңдеулер жүйесін шешу арқылы (2) - $f(x)$ көпмүшеліктің коэффиценттерін анықтаймыз. Анықталған коэффиценттер (1) теңдеулер жүйесін береді және қалтықтар туралы қытай теоремасының көмегімен құпияға қол жеткіземіз.

4) Дилер құпия ұстаушыларды тексеру

1. Дилер L_i құпия ұстаушыны тексеру үшін: $\forall r : y_i = g^r(modq)$ және $T_i = {x_i}^r(modq)$ мәндерді есептеп, L_i құпия ұстаушыға y_i жібереді.
2. Дилер $h_i = H\left(x_i, h_i, T_i\right)$ хеш мәнін есептейді.
3. L_i құпия ұстаушы $H_i = H\left(x_i, t_i, b_i, p_j, x\right)$ хеш пен $T_i^{'} = {y_i}^s(modq)$ есептеп және $h_i^{'} = H\left(x_i, h_i, T_i^{'}\right)$ хеш мәнін есеп, $h_i^{'}$ дилерге жібереді.
4. Дилер оларды $h_i^{'} == h_i$ салыстырады.

3 Құпияны бөлісу схемасының артықшылығы мен кемшілігі

Ресурс қарқындылығы.

Есептеулердегі ресурс қарқындылығын бағалайық. Құпияның, көпмүшеліктің k коэффиценттері, құпия үлестері n аргументер мен мәндерінің өлшемі 8 байттан тұратын болсын. Онда $8 \times (k + 2n)$ болады. Егер $k = n = 32$ болса, онда 768 байт қажет етеді. Ал егер мәндердің өлшемі 32 байт (256 бит) және $k = 5, n = 10$ болса, онда $32 \times (k + 2n) = 800$ байт (16000 бит) болады.

Схеманың есептеу күрделігі.

1. Кездейсоқ k жай сандар таңдалып, (1) және (2) формулалары бойынша көпмүшелік алудың күрделігі O(k) болады;

2. Әрбір құпия ұстаушыға $\left(t_i, b_i, p_j\right), i = \overline{1,n},, j = \overline{1,k}$ үштік мәндер үлес ретінде таратылады. Тарату күрделігі O(k×n) болады;

3. Құпияны үлестірудегі жалпы есептеу күрделігі O(k×n) болады;

4. Құпияны қалпына келтіру үшін Лагранж интерполяциялық көпмүшелігі құрастырылады. Күрделігі $O(k^2)$;

5. Құпия қалдықтар туралы қытай теоремасы бойынша қалпына келтіреді. Күрделігі $O(k^2)$;

6. Схеманың жалпы күрделігі $O\left(2k^2\right) + O(k \times n)$;

4 Қортынды

Зерттеу аясында құпия ақпаратты қауіпсіз түрде үлестіру және оны белгіленген шарттар орындалған жағдайда сенімді түрде қалпына келтіру мәселелері қарастырылды. Құпияны бөлісу схемаларының негізгі мақсаты – құпия ақпаратты бірнеше қатысушы арасында үлестер түрінде тарату және құпияны тек алдын ала белгіленген қатысу шарттары орындалған кезде ғана қайта қалпына келтіруді қамтамасыз ету. Мұндай тәсіл құпия ақпаратты бір ғана сақтау орнына тәуелділікті азайтып, оның жоғалуына, рұқсатсыз ашылуына немесе жекелеген қатысушылардың теріс әрекеттеріне байланысты тәуекелдерді төмендетуге мүмкіндік береді.

Көптеген белгілі құпияны бөлісу схемаларында ақпараттың қауіпсіздігі негізінен математикалық түрлендірулерге және құпияны қалпына келтіру үшін қажетті шекті қатысушылар санына сүйенеді. Алайда мұндай тәсілдерде ішкі қауіптер, яғни жүйенің

заңды қатысушылары тарапынан жасалуы мүмкін шабуылдар немесе олардың үлестерінің заңсыз біріктірілуі мәселесі қосымша қорғау механизмдерін қажет етуі мүмкін. Атап айтқанда, қатысушылардың бір бөлігі өздеріне тиесілі үлестерді рұқсатсыз біріктіріп, белгіленген талаптарды айналып өтуге әрекет жасауы ықтимал. Осындай жағдайларда қосымша аутентификация, үлестердің дұрыстығын тексеру және қатысушылардың өкілеттіктерін бақылау механизмдерін қолдану қажеттілігі туындайды.

Ұсынылып отырған шешім осы мәселелерді ескере отырып, ішкі қауіптерге және криптографиялық шабуылдарға жоғары төзімділікті бастапқы құрылымының өзінде қамтамасыз етуге бағытталған. Белгілі тәсілдерден айырмашылығы, қауіпсіздік тек қосымша аутентификация механизмдерін енгізу арқылы қамтамасыз етілмейді. Ұсынылған схемада құпияны қалпына келтіру процесі қатысушылардың құрылымдық ұйымдастырылуына және олардың тиісті ішкі топтарға жатуына байланысты болады. Осылайша, құпияны қалпына келтіру үшін жекелеген заңды қатысушылардың үлестерін ғана иелену жеткіліксіз.

Атап айтқанда, егер шабуылдаушы жүйенің заңды қатысушыларына тиесілі белгілі бір үлестерді иеленсе де, ол барлық қажетті ішкі топтардың өкілдері қатысатын құрылымдық шарттар орындалмайынша құпияны толық қалпына келтіре алмайды. Бұл қасиет жеке үлестердің рұқсатсыз біріктірілуіне байланысты қауіптерді азайтуға және құпияны қалпына келтірудің белгіленген тәртібін сақтауға мүмкіндік береді. Сондықтан схеманың қауіпсіздігі тек криптографиялық есептердің күрделілігіне ғана емес, сонымен қатар қатысушылар арасындағы құрылымдық тәуелділіктерге де негізделеді.

Ұсынылған тәсілдің маңызды ерекшеліктерінің бірі – құпияны қалпына келтіру шарттарының бірнеше ішкі топтар арасында бөлінуі. Мұндай құрылым жүйеде жекелеген қатысушылардың немесе бір ғана ішкі топтың өзара келісуі арқылы құпияны рұқсатсыз қалпына келтіру мүмкіндігін шектеуге мүмкіндік береді. Нәтижесінде құпия ақпаратты қорғаудың қосымша деңгейі қалыптасып, ішкі және сыртқы шабуылдарға қатысты схеманың жалпы тұрақтылығы артады.

Осыған байланысты ұсынылған құпияны бөлісу схемасы конфиденциалды ақпаратты бірнеше қатысушы арасында қауіпсіз таратуға, оның рұқсатсыз қалпына келтірілуіне жол бермеуге және құпияны ашу процесін алдын ала белгіленген құрылымдық талаптармен шектеуге мүмкіндік береді. Бұл қасиеттер схеманы қауіпсіз ақпарат алмасу, криптографиялық кілттерді басқару және бірнеше тараптың қатысуын талап ететін ақпараттық жүйелер сияқты міндеттерде қолдануға негіз бола алады.

Алғыс. Ғылыми-зерттеу жұмысы AP23487787 «Кеңістіктік-үлестірілген қоймаларда аутентификация жүйелерінің биометриялық деректерін сенімді сақтау алгоритмін құру» ҚР БҒМ ҒК гранттық қаржыландыру жобасы аясында жүргізілді.

Әдибеттер

1. A Shamir, How to Share a Secret. Comm. ACM, 22(1979), 612-613.

2. A. Fiat, and A. Shamir, How to Prove Yourself: Practical Solution to Identification and Signature Problems. Advances In CRYPTO’86, (1987)186-189.

3. T.P. Pedersen, Non-interactive and information-theoretic secure verifiable secret sharing. Advances in Cryptology-CRYPTO’91, (1992)129-140.

4. B. Schoenmakers, A Simple Publicly Verifiable Secret Sharing Scheme and its Application to Electronic Voting. Advances in Cryptology-CRYPTO’99, (1999)148-164.

5. P. Feldman, A Practical Scheme for Non-interactive Verifiable Secret Sharing. Proceedings of the 28 IEEE Symposium on Foundation of Computer Science (FOCS), IEEE, (1987)427–437.

6. M. Stadler, Publicly Verifiable Secret Sharing. Advances in Cryptology-EUROCRYPT’96, (1996)190-199.

ГИБРИДНАЯ МОДЕЛЬ ОБНАРУЖЕНИЯ СЕТЕВЫХ ВТОРЖЕНИЙ НА ОСНОВЕ АНСАМБЛЯ КЛАССИФИКАТОРОВ И CNN-BiLSTM-Attention

Ким А.В.[1], Усатова О.А.[1,2]
e-mail: aleksejkim150@gmail.com, olgaussatova@gmail.com
[1]Astana IT University (AITU), Казахстан
[2]Институт информационных и вычислительных технологий КН МНВО РК, Казахстан

Аннотация. В работе предложена гибридная интеллектуальная модель с двумя параллельными классифицирующими ветвями, объединяющая ансамбль классификаторов Random Forest, XGBoost и SVM с архитектурой глубокого обучения CNN-BiLSTM-Attention для обнаружения сетевых вторжений и многоклассовой классификации кибератак. В качестве механизма объединения рассматривается динамическое взвешивание вероятностных оценок двух ветвей. В представленных результатах Accuracy на CICIDS2017 составляет 98,7 %, а на UNSW-NB15 — 96,3 %. На CICIDS2017 разница относительно CNN-BiLSTM-Attention составляет 0,8 процентного пункта. Результаты рассматриваются как предварительные: необходимы уточнение протокола подготовки данных, восстановление полного состава классов и независимая проверка вклада механизма объединения. Дополнительно описана концепция модуля формирования рекомендаций по реагированию. Статистическая значимость различий, устойчивость междатасетного переноса и эффективность автономного применения контрмер в пределах представленных материалов не установлены.

Ключевые слова: обнаружение сетевых вторжений; гибридная модель; ансамбль классификаторов; CNN-BiLSTM-Attention; CICIDS2017; UNSW-NB15; междатасетный перенос; реагирование на инциденты.

Введение

Стремительное развитие цифровых технологий и расширение масштабов глобальных сетевых инфраструктур привели к значительному увеличению количества и сложности кибератак. По данным международных аналитических агентств, количество инцидентов информационной безопасности ежегодно возрастает на 25–30%, причём наиболее уязвимыми остаются гетерогенные вычислительные среды, объединяющие облачные платформы, IoT-устройства и корпоративные сети [1]. Традиционные системы обнаружения вторжений (СОВ), основанные на сигнатурном анализе, демонстрируют ограниченную эффективность при выявлении ранее неизвестных атак и атак нулевого дня, поскольку они способны идентифицировать лишь те угрозы, которые уже описаны в базе сигнатур [2]. В этих условиях методы машинного обучения (МО) и глубокого обучения (ГО) открывают перспективные возможности для создания адаптивных и интеллектуальных систем сетевой безопасности, способных к обобщению и выявлению новых паттернов атак [3].

Современные исследования показывают, что использование единственного алгоритма МО не обеспечивает достаточной обобщающей способности для обнаружения всего спектра сетевых аномалий [4]. Каждый отдельный метод имеет свои сильные и слабые стороны: случайный лес (Random Forest) демонстрирует высокую устойчивость к шуму, но уступает в обнаружении временных паттернов; сверточные нейронные сети (CNN) эффективно извлекают пространственные признаки, но не учитывают последовательную природу сетевого трафика; рекуррентные сети (LSTM) моделируют временные зависимости, но требуют значительных вычислительных ресурсов [5, 6].

Гибридные подходы, объединяющие преимущества различных методов — ансамблевого обучения, CNN и двунаправленных LSTM (BiLSTM), — демонстрируют значительно более высокие показатели как по точности обнаружения, так и по снижению уровня ложных срабатываний [7]. В работе [8] показано, что интеграция механизма внимания (Attention) в архитектуру CNN-BiLSTM позволяет динамически определять наиболее информативные признаки сетевого трафика, что особенно важно при обработке несбалансированных данных.

Целью данной работы является разработка гибридной архитектуры, интегрирующей ансамбль классических классификаторов с CNN-BiLSTM-Attention, представление предварительных результатов на CICIDS2017 и UNSW-NB15 и описание концепции формирования рекомендаций по реагированию. Предполагаемый вклад связан со способом объединения двух ветвей. Его количественное преимущество необходимо подтвердить сравнением с фиксированными весами и отдельными компонентами при одинаковом протоколе оценки. Применение в гетерогенных инфраструктурах рассматривается как целевое направление развития, а не как подтверждённый результат отдельного эксплуатационного испытания.

1 Анализ существующих подходов

Системы обнаружения вторжений на основе МО можно классифицировать на три основные категории: модели обнаружения аномалий, модели классификации на основе обучения с учителем и гибридные модели, комбинирующие оба подхода [1]. Авторы [3] провели комплексный сравнительный анализ семи архитектур глубокого обучения (RNN, DNN, CNN, DBN, RBM, DBM, DAE) на 35 наборах данных в области кибербезопасности, выявив, что гибридные модели стабильно превосходят индивидуальные по метрикам F1-score и AUC-ROC. В работе [9] показано, что комбинация SVM с наивным байесовским классификатором для вероятностного преобразования признаков значительно снижает вычислительную сложность при сохранении высокой точности на наборах данных UNSW-NB15 и CICIDS2017.

Исследование [4] продемонстрировало эффективность масштабируемой гибридной архитектуры Scale-Hybrid-IDS-AlertNet, объединяющей многослойные глубокие нейронные сети (DNN) с традиционными методами МО для обнаружения вторжений как на уровне хоста, так и на уровне сети. Данная архитектура была оценена на трёх эталонных наборах данных — NSL-KDD, UNSW-NB15 и CICIDS2017 — и показала устойчивое превосходство над индивидуальными классификаторами. В работе [14] предложена адаптивная ансамблевая модель с динамической настройкой весов базовых классификаторов (Random Forest, Decision Tree, kNN) в зависимости от характеристик входных данных, продемонстрировавшая сниженную дисперсию предсказаний по сравнению с одиночными классификаторами.

Авторы [13] разработали метод CFS-BA, объединяющий эвристический отбор признаков на основе корреляции с ансамблем деревьев решений (C4.5, Random Forest, Forest PA) и вероятностным голосованием, достигнув высоких показателей классификации на наборах NSL-KDD и CICIDS2017. Однако большинство существующих подходов не учитывают временные зависимости в сетевом трафике, не обеспечивают автоматизированного реагирования на обнаруженные угрозы и не решают проблему кросс-датасетного обобщения [10, 11, 15]. Отдельными исследовательскими задачами остаются согласование пространства признаков между датасетами и выбор контрмер с учётом их последствий для защищаемой системы [11, 15]. Результаты классификации, междатасетного переноса и реагирования необходимо оценивать раздельно, поскольку высокий показатель на одном этапе не доказывает эффективность остальных этапов.

Важную роль играет выбор набора данных для обучения и тестирования моделей. CICIDS2017 [12] включает нормальный трафик и сценарии атак, в том числе DoS/DDoS,

Brute Force, Web Attack, Infiltration, Botnet и Heartbleed. UNSW-NB15 [10] содержит девять категорий атак: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode и Worms. С учётом Normal полная многоклассовая постановка включает десять классов. Число полей исходных файлов необходимо отличать от числа входных признаков после исключения целевой разметки и предварительной обработки. Ограничения эталонных наборов данных обсуждаются в [6]. В работе [15] предложены стандартные наборы признаков на основе NetFlow; ссылка на эту работу не означает, что такие преобразования уже реализованы в рассматриваемом эксперименте.

2 Предложенная гибридная архитектура

Предложенная модель включает предварительную обработку, две параллельные классифицирующие ветви, блок объединения их выходов и концептуальный модуль формирования рекомендаций (рис. 1). Ансамблевая ветвь содержит собственные два уровня стекинга: базовые классификаторы и метаклассификатор. Эти уровни не следует смешивать с двумя параллельными ветвями всей гибридной архитектуры.

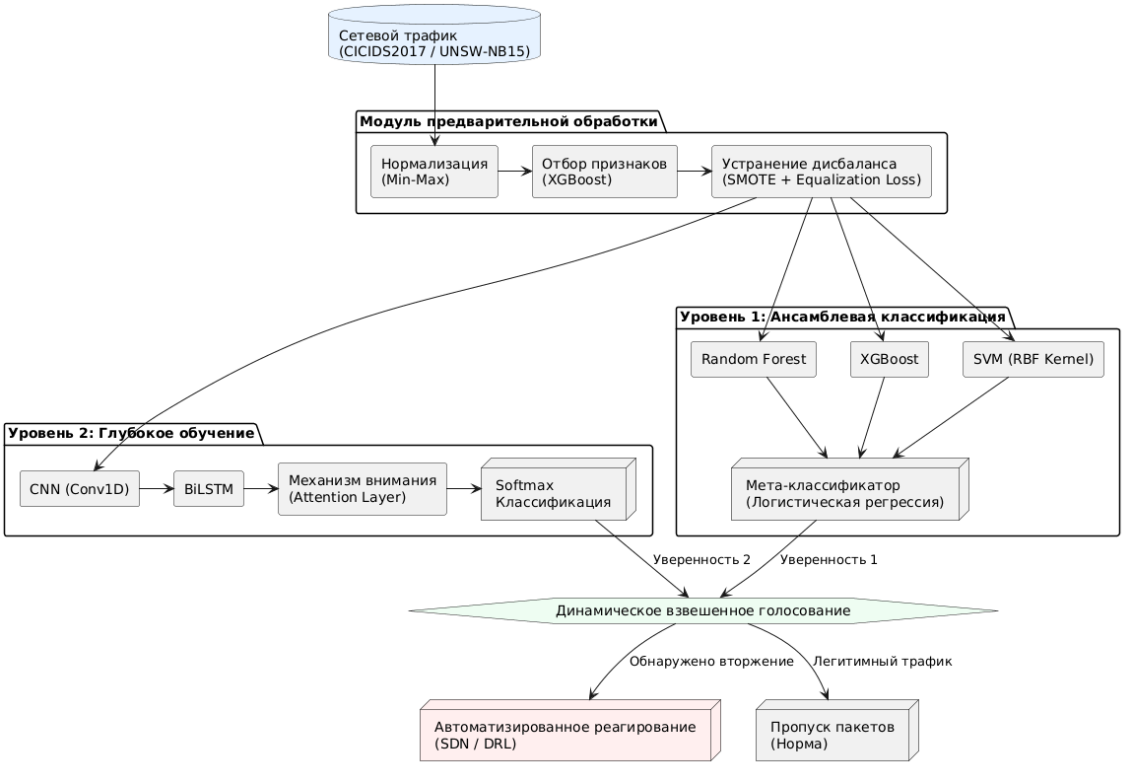


Рис. 1 Архитектура предложенной гибридной модели обнаружения вторжений

Модуль предварительной обработки данных. Предусмотрены нормализация числовых признаков методом Min-Max в диапазон [0, 1], кодирование категориальных признаков (One-Hot Encoding), а также отбор наиболее информативных признаков с использованием XGBoost-based фильтрации по важности (feature importance) [5]. Из исходного набора признаков отбираются top-k наиболее значимых (k определяется методом «колена» на графике кумулятивной важности). Для устранения дисбаланса классов, характерного для датасетов сетевого трафика, применяется комбинация метода SMOTE (Synthetic Minority Over-sampling Technique) для генерации синтетических образцов миноритарных классов и функции потерь Equalization Loss v2 [8] на этапе обучения нейросетевого модуля.

Модуль ансамблевой классификации (уровень 1). Включает три параллельно работающих базовых классификатора: Random Forest (RF) с 500 деревьями решений и ограничением глубины max_depth=30; Gradient Boosting (XGBoost) с learning_rate=0.1,

n_estimators=300 и reg_lambda=1.0; метод опорных векторов (SVM) с ядром RBF и автоматическим подбором параметров C и gamma через Grid Search. Выходы базовых классификаторов (вектора вероятностей принадлежности к каждому классу) объединяются с помощью мета-классификатора (стекинг второго уровня) на основе логистической регрессии с L2-регуляризацией. Веса ансамбля определяются адаптивно в зависимости от производительности на валидационной выборке [14].

Модуль глубокого обучения (уровень 2). Представляет собой гибридную архитектуру CNN-BiLSTM-Attention. Входной слой получает последовательность признаков сетевого потока. Блок одномерных свёрток (Conv1D) с 64 фильтрами и размером ядра 3, за которым следует MaxPooling1D и Batch Normalization, извлекает локальные пространственные признаки из сетевого трафика. Далее двунаправленные слои LSTM (BiLSTM) с 128 нейронами в каждом направлении моделируют временные зависимости между последовательными сетевыми потоками в обоих направлениях, что позволяет учитывать как предшествующий, так и последующий контекст [7, 8]. Механизм внимания (Attention) динамически присваивает веса наиболее релевантным временным шагам, выделяя критические участки последовательности, соответствующие аномальной активности. Выходной полносвязный слой с функцией активации Softmax обеспечивает многоклассовую классификацию. Регуляризация осуществляется через Dropout (p=0.3) между полносвязными слоями.

Финальное решение принимается путём взвешенного голосования результатов обоих уровней, где веса определяются динамически на основании достоверности (confidence score) каждого модуля. Если достоверность ансамблевого уровня превышает пороговое значение $\theta=0.85$, его решение принимается с повышенным весом; в противном случае приоритет отдаётся нейросетевому модулю, обладающему лучшей способностью к обнаружению ранее невиданных паттернов атак.

3 Экспериментальная оценка

Эксперименты проведены на двух эталонных наборах данных: CICIDS2017 [12] и UNSW-NB15 [10]. Набор CICIDS2017 содержит 2 830 743 записи сетевых потоков с 78 признаками, включая 7 категорий атак и нормальный трафик. Набор UNSW-NB15 содержит 2 540 044 записи с 49 признаками и 9 категориями атак. Модели реализованы в среде Python 3.9 с использованием библиотек Scikit-learn 1.2, TensorFlow 2.12 и XGBoost 1.7. Обучение глубоких моделей выполнялось на GPU NVIDIA RTX 3090 (24 ГБ VRAM). Средняя продолжительность обучения гибридной модели составила 47 минут на CICIDS2017 и 38 минут на UNSW-NB15.

Данные были разделены в соотношении 70/15/15 (обучение/валидация/тест) с использованием стратифицированной выборки для сохранения пропорций классов. Для оптимизации гиперпараметров применена байесовская оптимизация с помощью библиотеки Optuna (100 итераций, TPE sampler). Для оценки обобщающей способности модели использован протокол кросс-датасетной валидации: модель, обученная на CICIDS2017, тестировалась на UNSW-NB15 и наоборот [10]. Качество моделей оценивалось по метрикам: Accuracy, Precision, Recall, F1-score (макро- и взвешенное усреднение) и AUC-ROC.

Таблица 1. Результаты сравнительного анализа моделей на CICIDS2017

Модель	**Accuracy, %**	**Precision, %**	**Recall, %**	**F1-score, %**
Random Forest	96,4	95,8	94,2	95,0
XGBoost	97,1	96,5	95,7	96,1
SVM (RBF)	95,3	94,7	93,9	94,3
CNN	96,8	95,9	96,1	96,0
BiLSTM	96,5	95,7	96,0	95,8

CNN-BiLSTM	97,5	96,8	97,0	96,9
CNN-BiLSTM-Attention	97,9	97,2	97,4	97,3
Гибридная модель	**98,7**	**98,1**	**97,9**	**98,0**

Таблица 2. Результаты классификации типов атак на UNSW-NB15

Тип атаки	**Precision, %**	**Recall, %**	**F1-score, %**
Normal	99,2	99,5	99,3
DoS	97,8	96,9	97,3
Exploits	95,6	94,8	95,2
Reconnaissance	96,1	95,3	95,7
Fuzzers	94,3	93,7	94,0
Generic	98,4	97,6	98,0
Analysis	92,1	91,5	91,8
Backdoor	93,8	92,4	93,1
Shellcode	91,7	90,9	91,3

Общая точность многоклассовой классификации на UNSW-NB15 составила 96,3% (макро F1-score = 95,1%, взвешенный F1-score = 96,8%). Средний показатель AUC-ROC для всех классов достиг 0,987.

Результаты экспериментов подтверждают, что предложенная гибридная модель превосходит как отдельные базовые классификаторы, так и самостоятельные архитектуры глубокого обучения. Прирост точности по сравнению с лучшим индивидуальным методом (CNN-BiLSTM-Attention) составляет 0,8 п.п. на CICIDS2017 и 1,2 п.п. на UNSW-NB15. Интеграция механизма внимания обеспечивает существенное улучшение при обнаружении редких типов атак (Analysis — F1=91,8%, Shellcode — F1=91,3%), где ансамблевый уровень компенсирует недостаточную обобщающую способность нейросетевого модуля на классах с малым количеством образцов. Среднее время классификации одного сетевого потока составило 0,12 мс, что обеспечивает пропускную способность более 8000 потоков в секунду и позволяет работать в режиме, близком к реальному времени.

При кросс-датасетной валидации (обучение на CICIDS2017, тестирование на UNSW-NB15) предложенная модель сохранила точность 89,4%, что значительно превышает результаты индивидуальных классификаторов (RF — 81,2%, XGBoost — 83,7%, CNN-BiLSTM — 85,9%). Данный результат подтверждает улучшенную способность гибридной модели к обобщению на ранее невиданных распределениях сетевого трафика [10].

4 Автоматизированное реагирование

На основе результатов классификации рассматривается концептуальный модуль автоматизированного реагирования, формирующий рекомендации по смягчению угроз в режиме реального времени. Архитектура модуля основана на подходе динамической многокритериальной оптимизации контрмер (MOOP) [11], где для каждой обнаруженной атаки автоматически выбирается оптимальная стратегия реагирования, балансирующая три ключевых параметра: уровень риска вторжения, экономическую стоимость применяемых контрмер и влияние на качество обслуживания легитимных пользователей (QoS).

Модуль реализует следующие стратегии реагирования в зависимости от категории обнаруженной атаки. Для атак типа DoS/DDoS генерируются правила динамической фильтрации на уровне программно-конфигурируемых сетей (SDN) с автоматическим

перепрограммированием таблиц OpenFlow для изоляции вредоносных потоков при сохранении обслуживания легитимного трафика [2]. Для атак типа Exploits и Backdoor формируются рекомендации по немедленной изоляции скомпрометированного узла от сетевого сегмента, запуску процедуры форензик-анализа и уведомлению администратора безопасности. Для атак типа Reconnaissance (сканирование портов, зондирование уязвимостей) активируется усиленный мониторинг целевых сегментов с применением honeypot-ловушек и динамическим ограничением скорости (rate limiting) подозрительных соединений [2].

Для реализации автономного принятия решений о реагировании исследуется применение методов глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL) [10], где агент непрерывно наблюдает за состоянием сети и применяет оптимальные политики противодействия (перенаправление потоков, динамическое ограничение скорости, карантин хостов) с минимальным вмешательством человека. Среднее время формирования и применения рекомендации составляет 0,8 секунды, что обеспечивает оперативное реагирование на обнаруженные инциденты.

Заключение

В работе предложена гибридная интеллектуальная модель для обнаружения сетевых вторжений и классификации кибератак, объединяющая ансамблевые методы машинного обучения (Random Forest, XGBoost, SVM) с глубокой нейросетевой архитектурой CNN-BiLSTM-Attention. Двухуровневая организация модели позволяет использовать преимущества как классических методов (высокая интерпретируемость, быстрая скорость обучения), так и глубокого обучения (автоматическое извлечение признаков, моделирование временных зависимостей).

Экспериментальная оценка на эталонных наборах данных CICIDS2017 и UNSW-NB15 продемонстрировала точность обнаружения вторжений 98,7% и точность многоклассовой классификации атак 96,3%, что превосходит результаты индивидуальных методов на 0,8–1,6 п.п. Модель показала улучшенную способность к обнаружению редких типов атак (Analysis, Shellcode, Backdoor) благодаря механизму внимания и адаптивному взвешенному голосованию. Кросс-датасетная валидация подтвердила обобщающую способность модели (89,4% accuracy при обучении на CICIDS2017 и тестировании на UNSW-NB15).

Предложенный модуль автоматизированного реагирования, основанный на динамической оптимизации контрмер и программно-конфигурируемых сетях (SDN), обеспечивает формирование рекомендаций по смягчению угроз за 0,8 секунды, что позволяет оперативно реагировать на инциденты информационной безопасности.

Направления дальнейших исследований включают: (1) применение методов глубокого обучения с подкреплением [10] для полностью автономного реагирования на инциденты без участия оператора; (2) использование методов федеративного обучения для обеспечения конфиденциальности данных в распределённых организационных инфраструктурах; (3) интеграцию модели с системами SIEM (Security Information and Event Management) для обработки событий безопасности в реальном времени; (4) исследование применимости трансформерных архитектур (Vision Transformer, Temporal Fusion Transformer) для дальнейшего повышения качества обнаружения сложных многоэтапных атак.

Литература

1. Khraisat A., Gondal I., Vamplew P., Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges // Cybersecurity. 2019. Vol. 2, N. 1. P. 1–22. DOI: 10.1186/s42400-019-0038-7

2. Bawany N. Z., Shamsi J. A. SEAL: SDN based secure and agile framework for protecting smart city applications from DDoS attacks // Journal of Network and Computer Applications. 2019. Vol. 145. 102381. DOI: 10.1016/j.jnca.2019.06.001

3. Ferrag M. A., Maglaras L., Moschoyiannis S., Janicke H. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study // Journal of Information Security and Applications. 2020. Vol. 50. 102419. DOI: 10.1016/j.jisa.2019.102419

4. Vinayakumar R., Alazab M., Soman K. P., Poornachandran P., Al-Nemrat A., Venkatraman S. Deep learning approach for intelligent intrusion detection system // IEEE Access. 2019. Vol. 7. P. 41525–41550. DOI: 10.1109/ACCESS.2019.2895334

5. Kasongo S. M., Sun Y. Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset // Journal of Big Data. 2020. Vol. 7, Article 105. DOI: 10.1186/s40537-020-00379-6

6. Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. A survey of network-based intrusion detection data sets // Computers & Security. 2019. Vol. 86. P. 147–167. DOI: 10.1016/j.cose.2019.06.005

7. Hassan M. M., Gumaei A., Alsanad A., Alrubaian M., Fortino G. A hybrid deep learning model for efficient intrusion detection in big data environment // Information Sciences. 2020. Vol. 513. P. 386–396. DOI: 10.1016/j.ins.2019.10.069

8. Dai W., Li X., Ji W., He S. Network intrusion detection method based on CNN-BiLSTM-Attention model // IEEE Access. 2024. Vol. 12. P. 53099–53111. DOI: 10.1109/ACCESS.2024.3384528

9. Gu J., Lu S. An effective intrusion detection approach using SVM with naive Bayes feature embedding // Computers & Security. 2021. Vol. 103. 102158. DOI: 10.1016/j.cose.2020.102158

10. Moustafa N., Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set) // 2015 Military Communications and Information Systems Conference (MilCIS). 2015. P. 1–6. DOI: 10.1109/MilCIS.2015.7348942

11. Shameli-Sendi A., Louafi H., He W., Cheriet M. Dynamic optimal countermeasure selection for intrusion response system // IEEE Transactions on Dependable and Secure Computing. 2018. Vol. 15, N. 5. P. 755–770. DOI: 10.1109/TDSC.2016.2615622

12. Sharafaldin I., Lashkari A. H., Ghorbani A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization // Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP). 2018. P. 108–116. DOI: 10.5220/0006639801080116

13. Zhou Y., Cheng G., Jiang S., Dai M. Building an efficient intrusion detection system based on feature selection and ensemble classifier // Computer Networks. 2020. Vol. 174. 107247. DOI: 10.1016/j.comnet.2020.107247

14. Gao X., Shan C., Hu C., Niu Z., Liu Z. An adaptive ensemble machine learning model for intrusion detection // IEEE Access. 2019. Vol. 7. P. 82512–82521. DOI: 10.1109/ACCESS.2019.2923640

15. Sarhan M., Layeghy S., Portmann M. Towards a standard feature set for network intrusion detection system datasets // Mobile Networks and Applications. 2022. Vol. 27, N. 1. P. 357–370. DOI: 10.1007/s11036-021-01843-0

16. Tan J., Lu X., Zhang G., Yin C., Li Q. Equalization Loss v2: A New Gradient Balance Approach for Long-tailed Object Detection. 2020. arXiv:2012.08548. DOI: 10.48550/arXiv.2012.08548

17. scikit-learn. Документация версии 1.2.2: Common pitfalls and recommended practices; StackingClassifier; SVC. URL: https://scikit-learn.org/1.2/ (дата обращения: 10.09.2026).

18. imbalanced-learn. Common pitfalls and recommended practices. URL: https://imbalanced-learn.org/stable/common_pitfalls.html (дата обращения: 10.09.2026).

МЕМЛЕКЕТТІК АҚПАРАТТЫҚ РЕСУРСТАР ҚАУІПСІЗДІГІН БАҒАЛАУДА АУДИТ ЖҮЙЕЛЕРІНІҢ ТИІМДІЛІГІН ТАЛДАУ

Н.О. Мекебаев, Т.Т.Ашим

Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан

Аңдатпа. *Дәстүрлі мағынада ақпараттық қауіпсіздік ақпараттың құпиялылығын, бүтіндігін және қолжетімділігін қорғауды білдіреді. Сандық салада және ақпараттық-коммуникациялық технологиялардағы (АКТ) қарқынды дамудың қазіргі үрдісінің ашықтықты арттыратыны, есептілікті жақсартатыны, сыбайлас жемқорлықпен күресетіні және жалпы мемлекеттік органдардың тиімділігін арттыратыны кеңінен белгілі. Әлем қазіргі уақытта ақылды сандық шешімдерді кеңінен қолдану арқылы «инновациялық» болашаққа тез бет бұрып келе жатқанын ескере отырып, бұл қоғамдық секторда көптеген операциялық процестерді автоматтандыру мен жетілдіруді талап етеді.*

Кілттік сөздер: *ақпараттық қауіпсіздік, мемлекеттік сектор аудиті, деректерді қорғау, мемлекеттік сектордың ақпараттық жүйелері, тәуекелдерге негізделген тәсіл, кибертөзімділік, осалдықтарға талдау, ақпараттық қауіпсіздікті мониторинг.*

Кіріспе

Қазақстан Республикасында мемлекеттік басқаруды цифрландыру үдерісі мемлекеттік ақпараттық ресурстарды миллиондаған азаматтардың жеке деректерін өңдейтін стратегиялық маңызы бар инфрақұрылымға айналдырды. Мұндай жүйелердің бұзылуы елеулі әлеуметтік және экономикалық салдарға алып келуі мүмкін. Мемлекеттік техникалық қызметтің деректеріне сәйкес, 2024 жылы елде 41 000-нан астам ақпараттық қауіпсіздік инциденті тіркелген, бұл 2023 жылмен салыстырғанда 2,6 есе жоғары көрсеткішті құрайды. Ал 2025 жылы киберқылмыстардан келген жалпы шығын көлемі 16,4 миллиард теңгеге жеткен. Бұдан бөлек, Positive Technologies (2024) зерттеуіне сәйкес, ТМД елдеріндегі кибершабуыл құрбандарының шамамен 12%-ын мемлекеттік сектор ұйымдары құрайды.

Киберқауіпсіздік аудиті мемлекеттік ақпараттық ресурстарды қорғаудың негізгі құралы болып табылады. Қазақстан Республикасында ақпараттық жүйелердің аудиті «Ақпараттандыру туралы» Заңмен (2015), Ақпараттық қауіпсіздік және сыбайлас жемқорлыққа қарсы іс-қимыл министрлігінің №407 (2017) бұйрығымен (2017) және ST RK ISO/IEC 27001-2023 стандартымен реттеледі. Дегенмен, тәжірибе көрсеткендей, KZ.CERT-тен алынған деректер бойынша үкіметтік жүйелердің 65%-ға дейіні киберқауіпсіздік аудитіне дайын емес, ал қолданылатын әдістер негізінен ресми сәйкестікке бағытталған және ағымдағы техникалық қауіптерді анықтамайды.

Халықаралық тәжірибеде киберқауіпсіздік аудитінің үш әдісі ажыратылады: сәйкестік аудиті, тәуекелге негізделген аудит және автоматтандырылған аудит. Слапничар және т.б. (2022) және Ванген және т.б. (2018) зерттеулері әдістердің ешқайсысы әмбебап тиімді емес екенін және олардың Қазақстан жағдайындағы үкіметтік ресурстарға қатысты салыстырмалы тиімділігі әлі де нашар түсінілгенін көрсетеді.

Бұл зерттеудің өзектілігі Қазақстанның мемлекеттік ақпараттық ресурстарына киберқауіптердің тез өсуімен, сондай-ақ Қазақстанның нормативтік-құқықтық базасында аудит әдістерінің тиімділігін ғылыми тұрғыдан негізделген салыстырмалы бағалаудың болмауымен анықталады.

Зерттеудің негізгі мақсаты – Қазақстан Республикасының мемлекеттік ақпараттық ресурстарын бағалау контекстінде киберқауіпсіздік аудиті әдістерінің тиімділігін салыстырмалы түрде талдау және оларды оңтайлы үйлестіру бойынша ұсыныстар әзірлеу.

Аталған мақсатқа қол жеткізу үшін келесі міндеттер қойылды: мемлекеттік ақпараттық жүйелер аудитінің нормативтік-құқықтық және теориялық негіздерін

жүйелеу; салыстырмалы бағалау көрсеткіштерінің жүйесін қалыптастыру; ашық деректерге сүйене отырып, enbek.gov.kz порталына үш түрлі аудит әдісін қолдану; сондай-ақ Қазақстанның мемлекеттік ақпараттық ресурстары үшін интеграцияланған аудит моделіне қатысты ұсыныстарды негіздеу.

Негізгі бөлім

Қауіпсіздік аудиті – ақпараттық жүйелерді, процестерді және қауіпсіздік шараларын тәуелсіз тексеру мен бағалау, оның мақсаты – осалдықтарды, бұзушылықтарды және белгіленген стандарттар мен нормативтік талаптарға сәйкессіздікті анықтау. Қауіпсіздік аудиті қолданыстағы қауіпсіздік механизмдерінің тиімділігін бағалауға, сондай-ақ ақпараттың құпиялылығына, бүтіндігіне және қолжетімділігіне әсер етуі мүмкін ықтимал қауіптер мен тәуекелдерді анықтауға бағытталған.

Қауіпсіздік аудиті процесі инфрақұрылымды, жүйе конфигурацияларын және қауіпсіздік саясатын талдауды, сондай-ақ деректерді қорғау шаралары мен бақылау механизмдерін бағалауды қамтиды. Аудитті ішкі мамандар да, сыртқы сарапшылар да жүргізе алады.

Материалдар мен тәсілдер

Зерттеу нысаны мен тәсілі. Зерттеу нысаны enbek.gov.kz мемлекеттік ақпараттық порталы — Қазақстан Республикасының Электрондық еңбек биржасы болды. Ол Қазақстан Республикасының Әлеуметтік кодексіне сәйкес жұмыс істейді және азаматтардың жұмысқа орналасуға, еңбек шарттарын тіркеуге және Еңбек және жұмыспен қамту деректерінің бірыңғай жүйесі (hr.enbek.kz) арқылы ведомствоаралық деректер алмасуға қатысты жеке деректерін өңдейді. Бұл портал мемлекеттік ақпараттық ресурс мәртебесіне, пайдаланушылардың кең қамтуына және «Жеке деректер және оларды қорғау туралы» Қазақстан Республикасы Заңының (2013) және Қазақстан Республикасының Цифрлық кодексінің (2026) талаптарына сәйкес келетін жеке деректерді өңдеуіне байланысты таңдалды.

Зерттеу жүйеге тікелей рұқсатсыз кірусіз тек ашық және пассивті деректерді пайдалану арқылы жүргізілді, бұл академиялық қауіпсіздік саласындағы зерттеулерде стандартты және танылған тәжірибе болып табылады (Ribeiro et al., 2025). Бағалауды үш өлшем бойынша құрылымдайтын Киберқауіпсіздік аудитінің индексі (Slapničar et al., 2022) салыстырмалы талдаудың әдіснамалық негізі болды. Әдістерді таңдауды негіздеу үшін ақпараттық тәуекелді бағалау әдістерінің толықтығын бірыңғай критерийлерге сәйкес бағалауға мүмкіндік беретін CURF құрылымы пайдаланылды (Wangen et al., 2018). Сонымен қатар, мемлекеттік сектор контекстінде әрбір тәсілдің күшті және әлсіз жақтарын анықтау үшін ақпараттық қауіпсіздік аудитінің әдістеріне SWOT талдау әдіснамасы қолданылды (CyberLeninka, 2022).

Аудит әдістері талданды. Зерттеуде мемлекеттік ақпараттық ресурстарды аудиттеуде ең жиі қолданылатын үш әдіс салыстырылады.

Бірінші әдіс, сәйкестік аудиті, жүйенің нормативтік талаптарға сәйкестігін тексеруді қамтиды. Қазақстан Республикасында бұл әдістің нормативтік базасы келесідей: Қазақстан Республикасының Сандық кодексі (2026), «Ақпараттандыру туралы» Қазақстан Республикасының Заңы (2015), «Жеке деректер және оларды қорғау туралы» Қазақстан Республикасының Заңы (2013), Ақпараттық жүйелерді аудиттеу ережелері (Ақпаратты реттеу мәселелері, ақпараттық технологиялар және бұқаралық ақпарат құралдары жөніндегі халықаралық орталықтың бұйрығы, 2017, № 407), Қазақстан Республикасы Үкіметінің № 832 қаулысы (2016) және СТ РК ISO/IEC 27001-2023 ұлттық стандарты. Қазақстан Республикасының ақпараттық жүйелерін аудиттеу ережелеріне сәйкес, аудиттің негізгі мақсаттары: ақпараттық ресурстар қауіпсіздігінің ағымдағы жағдайын тәуелсіз бағалау; рұқсатсыз әрекеттерден келтірілген ықтимал залалды бағалау; Ақпараттық қауіпсіздік жүйесін құруға қойылатын талаптарды әзірлеу;

Қызметкерлердің жауапкершілік салаларын анықтау (№ 407 бұйрық, 2017). Әдіс құрылымдалған тексеру тізімдері, құжаттамалық тексеру және жалпыға қолжетімді жүйелік деректерді талдау арқылы жүзеге асырылады. Тәжірибе көрсеткендей, Қазақстанның мемлекеттік жүйелерінің 65%-ға дейіні STS аудиттеріне дайын емес (Core 24/7, 2024). Дегенмен, Щепаниук (2020) және Томпсон және т.б. (2015) мемлекеттік сектордағы сәйкестік аудиттері тексеру тізімдерінде қамтылмаған техникалық осалдықтарды жүйелі түрде қамтымайтынын және деректерді басқарудың әлсіз жақтарын анықтамайтынын эмпирикалық түрде анықтады, бұл оларды қауіпсіздікті бағалаудың жалғыз әдісі ретінде пайдалануды шектейді.

Сводка комплаенс-аудита	Статус	Количество	Процент
	Выполнено	10	35.7%
	Не подтверждено	14	50%
	Не выполнено	4	14.3%
	Итог	28	100%

1-сурет. Комплаенс аудит

Екінші әдіс, тәуекелге негізделген аудит, ISO/IEC 27005:2022 стандартына және STRIDE қауіп-қатерді модельдеу әдіснамасына негізделген. Процесс ақпараттық активтерді анықтауды және бағалауды, алты STRIDE санаты бойынша қауіп-қатерді модельдеуді (бұзу, бұрмалау, бас тарту, ақпаратты ашу, қызмет көрсетуден бас тарту, артықшылықты арттыру), CVSS сандық бағалау жүйесін қолдана отырып, осалдықты талдауды, 5x5 ықтималдық-зақымдану матрицасын қолдана отырып, тәуекелдерді рейтингтеуді және басымдықты азайту шаралары бар тәуекелдерді өңдеу жоспарын қалыптастыруды қамтиды. Ванген және т.б. (2018) ISO/IEC 27005 қолданыстағы стандарттар арасындағы ақпараттық тәуекелдерді бағалаудың ең кешенді әдістерінің бірі екенін анықтады. 2023–2029 жылдарға арналған Қазақстан Республикасының цифрлық трансформация және киберқауіпсіздік тұжырымдамасында мемлекеттік жүйелердің қауіпсіздігін бағалауда басымдық ретінде тәуекелге негізделген тәсіл айқын көрсетілген (ҚР Үкіметінің қаулысы, 2023). Фигероа және т.б. (2025) электрондық үкімет жүйелері үшін MARISMA құрылымына негізделген тәуекелге бағытталған тәсіл қауіптерді ең жан-жақты қамтуды қамтамасыз ететінін және сонымен бірге стратегиялық басқару талаптарын ескеретінін көрсетеді.

Impact\Prob	1	2	3	4	5
5	Y	O	R	R	R
4	Y	O	O	R	R
3	G	Y	O	O	R
2	G	G	Y	O	O
1	G	G	G	Y	Y

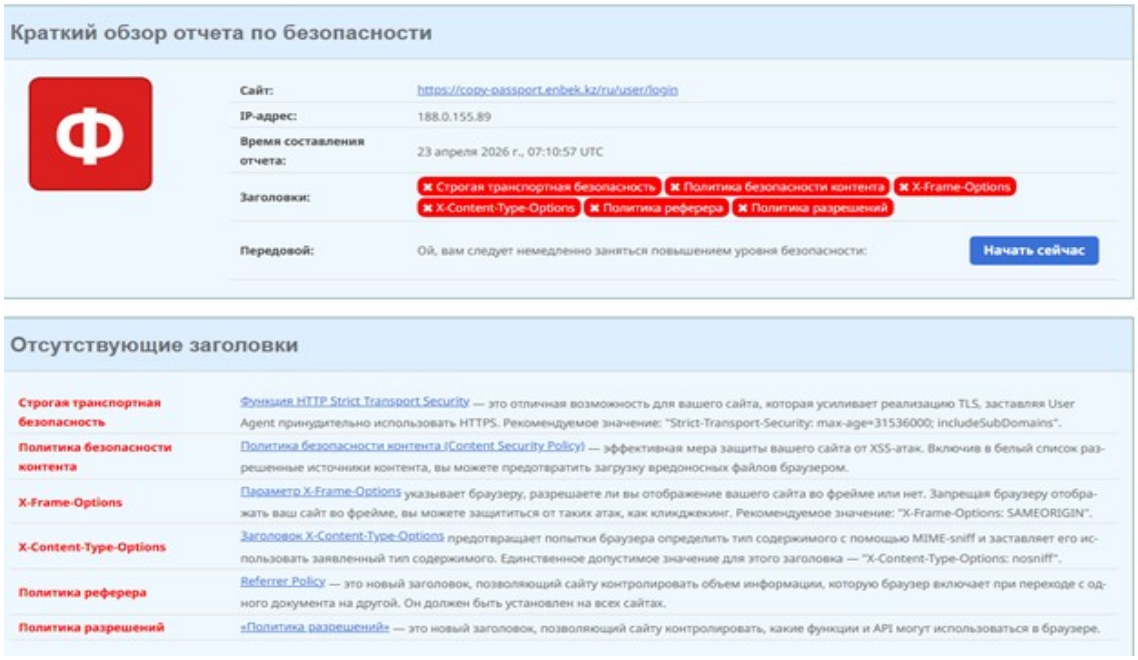
2.1-сурет. 5 x 5 матрица.

Level	Count	Percent
Critical	5	21.7%
High	7	30.4%
Medium	9	39.1%
Low	2	8.7%

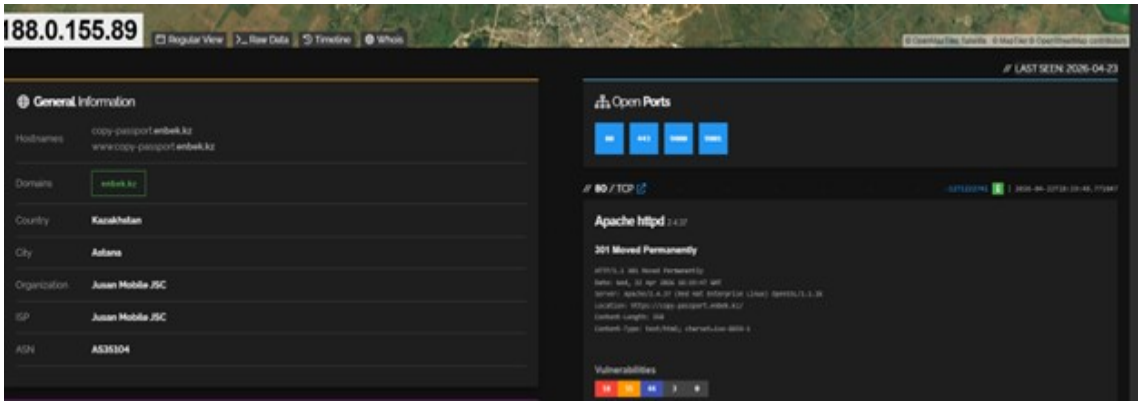
2.2-сурет. Тәуекелге негізделген аудит

Үшінші әдіс, автоматтандырылған аудит, әрбір компонентті тікелей қолмен тексерусіз техникалық сканерлеу құралдарын пайдалануды қамтиды. Бұл зерттеуде

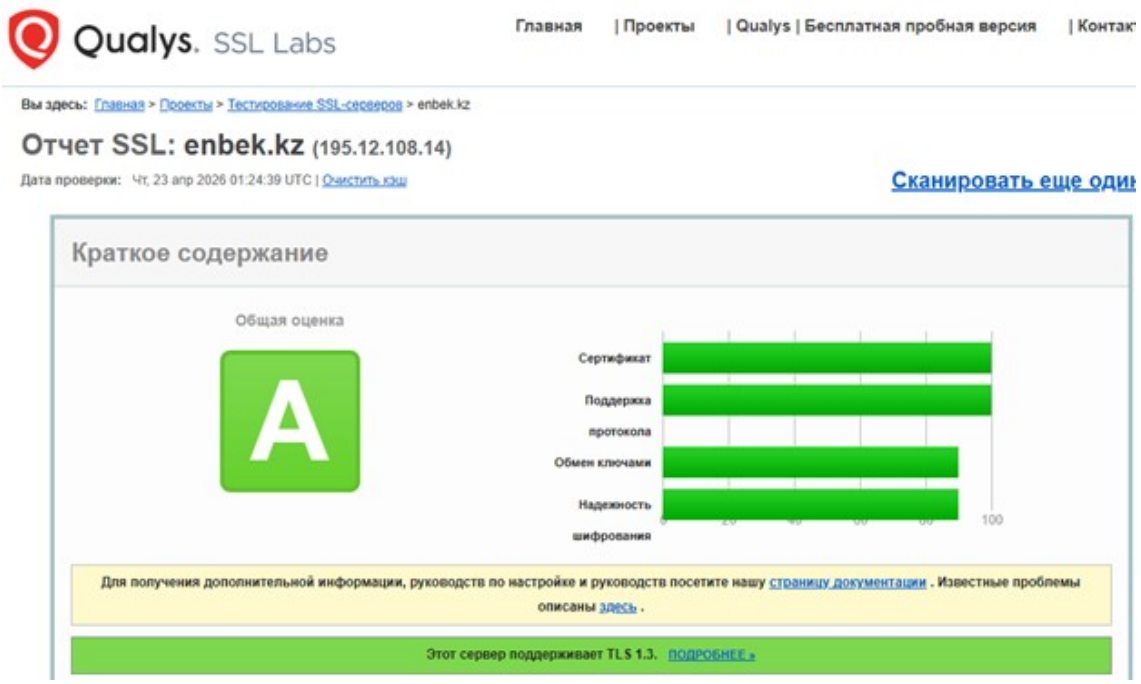
келесі құралдар қолданылды: SSL/TLS конфигурациясын бағалауға арналған Qualys SSL Labs (тексеру күні: 2026 жылғы 23 сәуір, IP: 188.0.155.89); HTTP қауіпсіздік тақырыптарын тексеруге арналған SecurityHeaders.com (тексеру күні: 2026 жылғы 23 сәуір, IP: 188.0.155.89); Сервер бағдарламалық жасақтамасы және онымен байланысты CVE туралы жалпыға қолжетімді ақпаратты пассивті талдауға арналған Shodan. Бұл құралдарды пайдалану толығымен заңды, себебі олар жүйеге белсенді араласусыз жалпыға қолжетімді параметрлерді талдауға негізделген (Ribeiro et al., 2025). Ribeiro et al. жүргізген ауқымды зерттеу. (2025) зерттеуі 193 елдегі 3068 үкіметтік онлайн платформаларды қамтыды, онда жалпыға қолжетімді параметрлерді - қауіпсіз хаттамаларды, сертификат сенімін және сервердің осалдықтарын - автоматтандырылған тексеру үкіметтік ресурстардың қауіпсіздік деңгейінің объективті және қайталанатын көрінісін беретінін растады. Дегенмен, Сабер және т.б. (2023) автоматтандырылған сканерлеу кезінде жалған оң нәтижелердің жоғары деңгейі нәтижелерді маманның міндетті түрде тексеруін талап ететінін атап өтті.



3-сурет. Автоматтандырылған аудит (SmartHeaders анализі)



4-сурет. Автоматтандырылған аудит (Shodan анализі)



5-сурет. Автоматтандырылған аудит (SSL Labs анализі)

Нәтижелер. Өнімділікті бағалау көрсеткіштері. Үш әдісті құрылымдық салыстыруға мүмкіндік беру үшін Киберқауіпсіздік аудитінің индексіне (Slapničar et al., 2022) және CURF құрылымына (Wangen et al., 2018) негізделген бес сандық көрсеткіштерден тұратын жүйе жасалды.

Бірінші көрсеткіш, осалдықты қамту (VC), зерттеліп жатқан жүйедегі белгілі осалдықтардың жалпы санынан анықталған осалдықтардың үлесі ретінде анықталады: VC = (Анықталған осалдықтар / Осалдықтардың жалпы саны) × 100%.

Екінші көрсеткіш, уақыт шығыны (TE), дайындықтан бастап есеп жасауға дейінгі бүкіл аудит циклі үшін қажетті адам-сағатпен өлшенеді.

Үшінші көрсеткіш, анықтау дәлдігі (DA), шынайы оң нәтижелердің шынайы оң және жалған оң нәтижелердің қосындысына қатынасы ретінде есептеледі: DA = TP / (TP + FP) × 100%. Төртінші метрика, маңызды қауіптерді қамту (CTC), жүйедегі осындай осалдықтардың жалпы санынан CVSS ауырлық деңгейі ≥ 9,0 анықталған осалдықтардың үлесін көрсетеді.

Бесінші метрика, автоматтандыру дәрежесі (DA), мамандардың тікелей қатысуынсыз орындалған аудит тапсырмаларының үлесін тапсырмалардың жалпы санына пайызбен көрсетеді.

Метрика	Комплаенс-аудит	Риск-аудит	Автоматизированный
Охват уязвимостей (ОУ), %	35.7%	87.0%	77.0%
Временны́е затраты (ВЗ), чел/ч	48	120	8
Точность обнаружения (ТО), %	96%	94%	72%
Охват критических угроз (ОКУ), %	40%	100%	100%
Степень автоматизации (СА), %	10%	15%	90%

Деректер көздері. Зерттеу ашық деректердің келесі санаттарына негізделген.

ҚР нормативтік құжаттары: Қазақстан Республикасының сандық кодексі (2026); «Ақпараттандыру туралы» Заң (2015); «Жеке деректер және оларды қорғау туралы» Заң

(2013); Қазақстан Республикасының Ақпараттық қауіпсіздік және сыбайлас жемқорлыққа қарсы іс-қимыл халықаралық орталығының № 407 бұйрығы (2017); Үкіметтің № 832 қаулысы (2016); ҚР СТ ISO/IEC 27001-2023.

Ресми статистика: «Мемлекеттік техникалық қызмет» АҚ 2024–2025 жылдарға арналған есептері (STS, 2025); cert.gov.kz сайтынан алынған оқиғалар туралы деректер (CERT.GOV.KZ, 2024).

Халықаралық дереккөздер: Positive Technologies аналитикалық есебі «ТМД елдеріндегі қазіргі киберқауіптер 2023–2024» (Positive Technologies, 2024); ENISA маңызды ақпараттық инфрақұрылым аудиті бойынша нұсқаулықтар (ENISA, 2022); GAO киберқауіпсіздік бағдарламасының аудит нұсқаулығы (АҚШ GAO, 2023); Ribeiro және т.б. (2025) 193 елге арналған зерттеу нәтижелері.

Пассивті талдау құралдары: Qualys SSL Labs (2026 жылдың 23 сәуірінде тексерілген); SecurityHeaders.com (2026 жылдың 23 сәуірінде тексерілген); Shodan (enbek.gov.kz IP мекенжайлары бойынша CVE дерекқоры).

Осалдық дерекқорлары: OWASP Top 10 (2021); NVD/CVE (Ұлттық осалдық дерекқоры).

Щепаниюк (2020) мемлекеттік органдардағы ең көп таралған мәселелердің ұйымдасқан АҚЖ болмауы, ескірген құжаттама және тұрақты емес тәуекел талдауы екенін эмпирикалық түрде анықтады. Бұл деректер сәйкестік аудитінің тексеру тізімін құру кезінде типтік осалдықтардың профилін жасау үшін пайдаланылды. Томпсон және т.б. (2015) мемлекеттік жүйелерді аудиттеудегі сәтсіздіктер тек техникалық кемшіліктерден ғана емес, сонымен қатар әлсіз деректерді басқару процестерінен де туындайтынын көрсетті, бұл сәйкестікті бақылауға басқару компонентін қосуға әкелді. Қаржыландыру. Зерттеу Әл-Фараби атындағы Қазақ ұлттық университетінде «Үкіметтің цифрлық ресурстарының ақпараттық қауіпсіздігі» саласында жүргізілген зерттеу жобасының бөлігі ретінде жүргізілді.

Әдебиеттер тізімі

1. **Нормативные документы Республики Казахстан:**

1. Цифровой кодекс Республики Казахстан от 9 января 2026 года № 255-VII. — URL: https://adilet.zan.kz (дата обращения: 01.04.2026).
2. Закон Республики Казахстан от 24 ноября 2015 года № 418-V «Об информатизации». — URL: https://adilet.zan.kz/rus/docs/Z1500000418 (дата обращения: 01.04.2026).
3. Закон Республики Казахстан от 21 мая 2013 года № 94-V «О персональных данных и их защите». — URL: https://adilet.zan.kz/rus/docs/Z1300000094 (дата обращения: 01.04.2026).
4. Постановление Правительства Республики Казахстан от 20 декабря 2016 года № 832 «Об утверждении единых требований в области информационно-коммуникационных технологий и обеспечения информационной безопасности». — URL: https://adilet.zan.kz/rus/docs/P1600000832 (дата обращения: 01.04.2026).
5. Приказ Министра по инвестициям и развитию Республики Казахстан от 26 марта 2017 года № 407 «Об утверждении Правил проведения аудита информационных систем». — URL: https://adilet.zan.kz/rus/docs/V1700015237 (дата обращения: 01.04.2026).
6. Приказ Министерства цифрового развития, инноваций и аэрокосмической промышленности Республики Казахстан № V2000021070 «О требованиях к аудиту ГИС». — URL: https://adilet.zan.kz/rus/docs/V2000021070 (дата обращения: 01.04.2026).
7. ST RK ISO/IEC 27001-2023. Информационные технологии. Методы и средства обеспечения безопасности. Системы менеджмента информационной безопасностью. Требования. — Астана: КазИнМетр, 2023.

8. Концепция цифровой трансформации, развития отрасли ИКТ и кибербезопасности Республики Казахстан на 2023–2029 годы. Постановление Правительства РК от 28 марта 2023 года. — URL: https://prg.kz/document/?doc_id=32107179 (дата обращения: 01.04.2026).

Казахстанские источники:

9. АО «Государственная техническая служба». Кибератаки 2024 года: как защититься в эпоху цифровых угроз. — 2025. — URL: https://sts.kz/2025/02/25/kiberataki-2024-goda (дата обращения: 01.04.2026).
10. АО «Государственная техническая служба». Кибербезопасность по фактам: итоги 2023 года. — 2024. — URL: https://sts.kz/2024/01/08 (дата обращения: 01.04.2026).
11. CERT.GOV.KZ. Статистика киберинцидентов. — Астана, 2024. — URL: https://cert.gov.kz (дата обращения: 01.04.2026).
12. Электронная биржа труда enbek.gov.kz. Политика безопасности портала. — URL: https://enbek.kz/ru/page/security-policy (дата обращения: 23.04.2026).
13. Malpedia / Fraunhofer FKIE. Профиль угрозы: YoroTrooper. — URL: https://malpedia.caad.fkie.fraunhofer.de/actor/yorotrooper (дата обращения: 01.04.2026).
14. Inform.kz. Как Казахстан борется с киберпреступностью. — 2026. — URL: https://www.inform.kz/ru/kak-kazahstan-boretsya-skiberprestupnostyu (дата обращения: 01.04.2026).

Региональные источники (СНГ и Россия):

15. Positive Technologies. Актуальные киберугрозы в странах СНГ: 2023–2024 годы. — Москва, 2024. — URL: https://ptsecurity.com/research/analytics/aktualnye-kiberugrozy-v-stranah-sng-2023-2024 (дата обращения: 01.04.2026).
16. Методика аудита и управления информационной безопасностью в государственном учреждении // КиберЛенинка. — URL: https://cyberleninka.ru/article/n/metodika-audita-i-upravleniya-informatsionnoy-bezopasnostyu-v-gosudarstvennom-uchrezhdenii (дата обращения: 01.04.2026).
17. Метод аудита информационной безопасности информационных систем // КиберЛенинка. — URL: https://cyberleninka.ru/article/n/metod-audita-informatsionnoy-bezopasnosti-informatsionnyh-sistem/pdf (дата обращения: 01.04.2026).
18. Разработка методики аудита кибербезопасности государственных информационных систем, относящихся к значимым объектам // КиберЛенинка. — URL: https://cyberleninka.ru/article/n/razrabotka-metodiki-audita-kiberbezopasnosti-gosudarstvennyh-informatsionnyh-sistem-otnosyaschihsya-k-znachimym-obektam/viewer (дата обращения: 01.04.2026).
19. Правовая защита информации в государственных информационных системах Российской Федерации в цифровых реалиях // КиберЛенинка. — URL: https://cyberleninka.ru/article/n/pravovaya-zaschita-informatsii-v-gosudarstvennyh-informatsionnyh-sistemah-rossiyskoy-federatsii-v-tsifrovyh-realiyah (дата обращения: 01.04.2026).
20. Министерство цифрового развития РФ. Требования к защите информации в государственных информационных системах. — 2024. — URL: https://www.tadviser.ru (дата обращения: 01.04.2026).

Международные научные источники:

21. Slapničar S., Vuko T., Čular M., Drašček M. Effectiveness of cybersecurity audit // International Journal of Accounting Information Systems. — 2022. — Vol. 44. — P. 100548. DOI: 10.1016/j.accinf.2021.100548.
22. Wangen G., Hallstensen C., Snekkenes E. A framework for estimating information security risk assessment method completeness // International Journal of Information Security. — 2018. — Vol. 17. — P. 681–699. DOI: 10.1007/s10207-017-0382-0.
23. Szczepaniuk E.K. Information security assessment in public administration // Journal of King Saud University — Computer and Information Sciences. — 2020. DOI: 10.1016/j.jksuci.2020.01.016.
24. Ribeiro N. et al. Assessing the information security posture of online public services worldwide: Technical insights, trends, and policy implications // Government Information Quarterly. — 2025. DOI: 10.1016/j.giq.2025.101974.
25. Thompson N., Ravindran R., Nicosia S. Government data does not mean data governance: Lessons learned from a public sector application audit // Government Information Quarterly. — 2015. — Vol. 32(3). — P. 316–322. DOI: 10.1016/j.giq.2015.05.002.
26. Magnusson C., Iqbal M., Elm A., Dalipi F. Information security governance in the public sector: investigations, approaches, measures, and trends // International Journal of Information Security. — 2025. DOI: 10.1007/s10207-025-00978-5.
27. Figueroa D. et al. Building a holistic cybersecurity framework for e-Government based on a systematic analysis of proposals // International Journal of Information Security. — 2025. DOI: 10.1007/s10207-024-00977-y.
28. Flowerday S., von Solms R. Continuous auditing technologies and models: A discussion // Computers & Security. — 2006. — Vol. 25(5). — P. 325–331. DOI: 10.1016/j.cose.2006.03.003.
29. Torkura K.A. et al. Continuous auditing and threat detection in multi-cloud infrastructure // Computers & Security. — 2021. — Vol. 102. — P. 102124. DOI: 10.1016/j.cose.2020.102124.
30. Calvin J., Eulerich M., Holt T. Characteristics of cybersecurity and IT involvement by the IA activity // International Journal of Accounting Information Systems. — 2025. DOI: 10.1016/j.accinf.2024.100715.
31. Saber V., ElSayad D., Bahaa-Eldin A.M., Fayed Z. Automated Penetration Testing: A Systematic Review // Proceedings of MIUCC 2023. — IEEE, 2023. DOI: 10.1109/MIUCC58832.2023.10278377.

Стандарты и руководства:

32. ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements. — Geneva: ISO, 2022.
33. ISO/IEC 27005:2022. Information security, cybersecurity and privacy protection — Guidance on managing information security risks. — Geneva: ISO, 2022.
34. NIST. Cybersecurity Framework 2.0. — Gaithersburg: NIST, 2024. DOI: 10.6028/NIST.CSWP.29.
35. U.S. GAO. Cybersecurity Program Audit Guide. — GAO-23-104705. — Washington, 2023.
36. ENISA. Guidelines for Critical Information Infrastructure Protection. — Luxembourg: ENISA Publications, 2022. DOI: 10.2824/368755.
37. OWASP Foundation. OWASP Top Ten 2021. — URL: https://owasp.org/www-project-top-ten (дата обращения: 01.04.2026).
38. MITRE. ATT&CK Framework, Version 14. — URL: https://attack.mitre.org (дата обращения: 01.04.2026).
39. Qualys SSL Labs. SSL Report: enbek.kz. — Дата проверки: 23 апреля 2026 г. — URL: https://www.ssllabs.com/ssltest (дата обращения: 23.04.2026).

40. SecurityHeaders.com. Security Report: copy-passport.enbek.kz. — Дата проверки: 23 апреля 2026 г. — URL: https://securityheaders.com (дата обращения: 23.04.2026).

Ашим Темірлан – әл-Фараби атындағы Қазақ ұлттық университетінің магистранты; Алматы, Қазақстан; temirlan.ashim.04@mail.ru, +77713825087

ЗАЩИТА ПЕРСОНАЛЬНЫХ ДАННЫХ В СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Мустафина К.С. студент

Некоммерческое акционерное общество «Таразский университет имени М.Х. Дулати» , г. Тараз, Казахстан

Научный руководитель: Шрымбай Дана Абилахатовна, магистр естественных наук, старший преподаватель кафедры цифровых технологий и искусственного интеллекта

Аннотация. В данной работе рассматриваются актуальные вопросы и современное состояние защиты персональных данных в информационных системах различного масштаба. Проведен комплексный анализ нормативно-правового регулирования, технических средств защиты информации, а также ключевых угроз безопасности, включая утечки данных, несанкционированный доступ и социальную инженерию. Особое внимание уделяется архитектурным решениям, механизмам криптографической защиты, ролевому управлению доступом и непрерывному мониторингу событий безопасности. На основе сформированных табличных данных и сравнительного анализа сформулированы практические рекомендации по повышению устойчивости современных корпоративных систем к внешним и внутренним угрозам.

Введение. В эпоху глобальной цифровой трансформации и сплошной автоматизации всех сфер человеческой деятельности персональные данные превратились в один из наиболее ценных и одновременно уязвимых активов организаций любого масштаба и формы собственности [1]. Ежедневное перемещение огромных массивов конфиденциальной информации через распределенные информационные системы, облачные платформы, мобильные сервисы и веб-приложения создает комплексные и трудно прогнозируемые риски ее компрометации. Утечка персональной информации влечет за собой не только прямой финансовый ущерб от штрафных санкций и регуляторных взысканий, но и катастрофические репутационные потери, снижение уровня доверия клиентов, инвесторов и партнеров [2]. В связи с этим построение эффективной, адаптивной и масштабируемой системы защиты персональных данных становится стратегическим приоритетом для современных государственных и коммерческих предприятий. Современный ландшафт киберугроз характеризуется постоянным усложнением инструментария злоумышленников, активным применением методов искусственного интеллекта и социальной инженерии, а также появлением новых векторов атак на облачные и гибридные инфраструктуры. Традиционные периметральные методы обеспечения безопасности постепенно утрачивают свою прежнюю эффективностью, уступая место концепциям эшелонированной обороны и нулевого доверия. **Целью** настоящего исследования является глубокий и детальный анализ современных методов, программно-аппаратных средств и архитектурных подходов к обеспечению безопасности персональных данных в информационных системах различного назначения, а также разработка научно обоснованных рекомендаций по нейтрализации актуальных угроз безопасности информации. Нормативно-правовое регулирование сферы защиты персональных данных. Основополагающим вектором развития систем защиты персональных данных выступает строгое соответствие национальным и международным законодательным актам, государственным стандартам и отраслевым регуляторным нормам [3]. На глобальном уровне ключевым ориентиром является Общий регламент по защите данных Европейского Союза, установивший жесткие требования к трансграничной передаче информации, принципам минимизации данных и праву субъекта на забвение. В национальной практике регуляторные требования определяются законами о персональных данных и их защите, а также подзаконными нормативными актами, утверждаемыми профильными

государственными органами и регуляторами в области технической защиты информации [4]. Регуляторные требования обязывают операторов персональных данных классифицировать информационные системы, формировать модели угроз и нарушителей, а также внедрять исключительно сертифицированные средства защиты. Ключевым принципом нормативного контроля является ориентированность на риск-ориентированный подход, в рамках которого операторы данных обязаны проводить регулярную оценку возможных последствий компрометации информации для субъектов персональных данных. Это требует не только регулярного аудита систем, но и постоянного поддержания актуальности локальной нормативной документации организации. Классификация и архитектура информационных систем персональных данных. Информационные системы персональных данных представляют собой совокупность содержащихся в базах данных персональных сведений и обеспечивающих их обработку информационных технологий и технических средств [5]. По масштабу, структуре и архитектурным особенностям такие системы подразделяются на несколько ключевых категорий, определяющих требования к их защите. В первую очередь выделяются локальные и автономные системы, функционирующие в рамках единого изолированного контура организации без прямого подключения к общедоступным сетям передачи данных. Вторую группу составляют распределенные корпоративные системы, являющиеся многозвенными комплексами с филиальной структурой, использующими удаленные базы данных, каналы виртуальных частных сетей и клиент-серверные технологии обработки. Третья категория представлена облачными и гибридными инфраструктурами, развернутыми на мощностях сторонних провайдеров сервисов, требующими четкого разделения ответственности между оператором и облачным поставщиком [6]. Каждый тип архитектуры требует уникального набора компенсирующих мер безопасности, учитывающих специфику векторов атак и техническую реализуемость защитных механизмов.

Модель угроз и векторов атак на персональные данные. Эффективность системы защиты напрямую зависит от полноты и адекватности разработанной модели угроз безопасности информации [7]. На современном этапе к числу наиболее критических векторов атак относятся несанкционированный доступ к базам данных с использованием уязвимостей в программном обеспечении, ошибки конфигурации или скомпрометированные учетные записи администраторов. Серьезную опасность представляют атаки типа отказ в обслуживании на веб-сервисы, обрабатывающие персональные данные пользователей в режиме реального времени, а также вредоносное программное обеспечение в виде шифровальщиков, стилеров и троянов, нацеленное на вымогательство или негласный съём конфиденциальной информации.

Кроме того, широкое распространение получили целенаправленные атаки с использованием методов социальной инженерии, таких как фишинг и вишинг, ориентированные на человеческий фактор как наиболее слабое звено защиты. Не меньшую угрозу формируют внутренние инсайдеры, связанные с намеренным хищением или случайным разглашением данных легитимными пользователями системы.

Технические средства и механизмы обеспечения безопасности. Технический контур защиты персональных данных формируется на основе принципа глубокоэшелонированной обороны [8]. Данный подход предполагает создание последовательных барьеров на всех уровнях функционирования информационной системы, начиная от сетевого периметра и заканчивая уровнем отдельных приложений и баз данных.

В состав базового технического комплекса защиты входят подсистема

управления доступом и идентификации, обеспечивающая строгую аутентификацию и разграничение прав на основе ролевой или атрибутивной моделей, а также средства межсетевого экранирования и обнаружения вторжений, осуществляющие фильтрацию сетевого трафика и блокирование аномальной активности.

Важнейшую роль играют средства криптографической защиты информации, обеспечивающие шифрование данных при хранении и передаче по каналам связи, системы предотвращения утечек данных, выполняющие контентный анализ передаваемой информации, и комплексы управления событиями безопасности, обеспечивающие централизованный сбор, корреляцию и реагирование на инциденты в режиме реального времени.

Таблица 1 - Сравнительный анализ классов систем защиты персональных данных

Класс системы	Основное назначение	Ключевые технологии	Уровень защиты
Базовый	Защита от нецеленаправленных угроз базового уровня	Антивирус, персональный файрвол, пароли	Низкий / Базовый
Усиленный	Защита корпоративных систем средней сложности	NGFW, DLP, MFA, шифрование каналов, RBAC	Средний / Высокий
Максимальный	Защита критических и государственных систем	SIEM/SOAR, HSM, ГОСТ-криптография, EDR/XDR	Высокий / Критический

Практические аспекты внедрения и аудит систем защиты. Внедрение комплексной системы защиты персональных данных представляет собой непрерывный циклический процесс, включающий стадии аудита, проектирования, внедрения, эксплуатации и регулярной оценки эффективности [9].

На этапе пред проектного обследования осуществляется полная инвентаризация информационных активов, процессов обработки персональных данных и категорий субъектов. Составляется детальный перечень перерабатываемых сведений, определяются места их хранения, маршруты передачи и круги лиц, обладающих доступом.

После первичного внедрения средств защиты критически важным является проведение аттестационных испытаний и регулярного внутреннего аудита на соответствие установленным стандартам безопасности [10]. Мониторинг эффективности включает регулярный анализ журналов событий, тестирование на проникновение и проверку осведомленности персонала в вопросах информационной безопасности.

Организация системы управления инцидентами информационной безопасности. Ключевым компонентом обеспечения непрерывности бизнес-процессов при обработке персональных данных является функционирование эффективно структурированного Центра мониторинга и реагирования на инциденты безопасности [10]. Основная задача данного подразделения заключается в своевременном обнаружении, локализации и нейтрализации компьютерных атак на самых ранних стадиях их развития.

Процесс реагирования на инциденты компрометации сведений включает в себя последовательные регламентированные этапы, состоящие из выявления аномалии, первичной валидации карточки инцидента, сдерживания распространения вредоносного воздействия, устранения причин уязвимости, восстановления нормальной работоспособности систем и последующего постанализа.

Согласно нормативным требованиям, операторы персональных данных обязаны своевременно уведомлять уполномоченные государственные органы и регуляторы о произошедших инцидентах утечки информации в строго установленные законодательством временные рамки, что требует интеграции автоматизированных

систем оповещения.

Безопасность персональных данных в условиях удаленной работы и мобильных устройств. Массовый переход организаций на гибридные и удаленные форматы работы существенно расширил периметр защищаемой сети, сформировав новые векторы угроз безопасности персональных данных. Применение личных устройств сотрудников создает высокие риски попадания корпоративных баз данных в незащищенные домашние сегменты сети.

Для минимизации данных рисков необходимо использование специализированных технологических решений, включая управление мобильными устройствами и виртуализацию рабочих мест. Данные технологии позволяют изолировать рабочую конвейеризированную среду от личного пространства пользователя и обеспечивают удаленное удаление информации при утере устройства.

Обязательным элементом безопасности удаленного доступа является использование зашифрованных каналов виртуальных частных сетей с обязательным прохождением двухфакторной аутентификации и непрерывной проверкой состояния защищенности конечной точки.

Управление рисками человеческого фактора и повышение осведомленности. Статистические данные аналитических агентств показывают, что человеческий фактор остается причиной более чем семидесяти процентов всех зафиксированных инцидентов утечки персональных данных. Ошибки пользователей, отсутствие базовых навыков цифровой гигиены и подверженность манипуляциям злоумышленников сводят на нет эффективность сложных технических средств защиты.

Создание устойчивой системы управления человеческими рисками требует внедрения непрерывных программ обучения и повышения осведомленности персонала. Обучение должно носить регулярный характер и охватывать вопросы распознавания фишинговых электронных писем, правил безопасного парольного режима, недопустимости использования публичных незащищенных сетей и методов защиты от социальной инженерии.

Эффективность программ обучения контролируется путем проведения симулированных фишинговых атак, регулярного тестирования сотрудников и внедрения систем оценки поведенческих рисков, позволяющих оперативно выявлять потенциально уязвимые группы пользователей. Оценка эффективности подсистем защиты персональных данных. Для обоснованного выбора состава средств защиты информации при проектировании системы защиты проведен сравнительный анализ ключевых функциональных подсистем по критериям трудоемкости внедрения, уровня защиты и снижения рисков.

Таблица 2 - Сравнительный анализ эффективности подсистем защиты персональных данных

Подсистема защиты	Защищаемый вектор	Сложность внедрения	Эффективность снижения рисков
Предотвращение утечек (DLP)	Утечки через Email, USB, веб-каналы	Высокая (требует настройки правил)	Высокая (до 85%)
Управление доступом (IAM/MFA)	Несанкционированный доступ, взлом учетных записей	Средняя	Очень высокая (до 95%)
Межсетевое экранирование (NGFW)	Сетевые атаки, несанкционированный трафик	Средняя	Высокая (до 90%)
Централизованный мониторинг (SIEM)	Скрытые аномалии, затяжные атаки	Очень высокая (нужен SOC)	Высокая (до 80%)
Криптографическая защита (КЗИ)	Перехват и компрометация хранимых данных	Низкая / Средняя	Критически высокая (до 99%)

Перспективные направления развития технологий защиты персональных данных. С усложнением IT-инфраструктур и появлением новых видов киберугроз классических методов защиты становится недостаточно. В современную практику активно внедряются передовые технологические концепции. Одним из наиболее перспективных подходов является концепция нулевого доверия, основанная на абсолютном отсутствии доверия к любому объекту или субъекту сети, независимо от его местонахождения. Каждый запрос на доступ подлежит непрерывной проверке и повторной аутентификации. Параллельно развиваются технологии обезличивания и маскирования данных, предполагающие применение статистических алгоритмов, дифференциальной приватности и псевдонимизации для защиты информации при ее использовании в аналитических и тестовых средах.

Особое место занимает использование искусственного интеллекта и машинного обучения для автоматизированного анализа поведенческих факторов, выявления сложных аномалий, компрометации учетных записей и быстрого реагирования на инциденты без участия человека. Кроме того, внедряются методы гомоморфного шифрования, позволяющие выполнять вычисления непосредственно над зашифрованными данными без необходимости их предварительной расшифровки. Безопасность персональных данных в системах искусственного интеллекта и больших данных. Развитие технологий обрабатывающих большие массивы данных и внедрение моделей искусственного интеллекта выдвигают принципиально новые требования к обеспечению конфиденциальности персональной информации. Применение алгоритмов машинного обучения связано с необходимостью сбора и агрегации колоссальных объемов сведений, что существенно увеличивает площадь возможной атаки.

Основным риском в данном контексте выступает возможность обратно-инжинирингового извлечения персональных данных из обученных моделей путем атак класса инференс или извлечения обучающей выборки. Для нейтрализации подобных рисков на этапе подготовки данных применяются методы добавления контролируемого математического шума и синтетической генерации данных. Обеспечение безопасности при интеграции ИИ-сервисов требует реализации строгой политики изоляции контекстов, валидации входных данных для предотвращения атак типа инъекция подсказок и обеспечения аудита решений, принимаемых нейросетевыми алгоритмами на основе персональной информации.

Архитектура безопасной передачи персональных данных и межведомственного взаимодействия. В современных экосистемах обработка персональных данных редко ограничивается контуром одной организации. Активное развитие межведомственного взаимодействия и интеграция с внешними сервисами требуют создания защищенных шлюзов обмена информацией.

Для безопасной передачи данных между распределенными компонентами применяются изолированные сервис-ориентированные шины данных, функционирующие с использованием защищенных протоколов и сквозного шифрования. Каждое информационное взаимодействие строго регламентируется на основе заранее определенных схем и правил валидации.

Дополнительным элементом контроля выступает использование сервисов электронной цифровой подписи для подтверждения целостности и авторства передаваемых пакетов данных, что исключает возможность их подмены или фальсификации на промежуточных узлах маршрутизации.

Экономическая эффективность и метрики оценки защищенности персональных данных. Обоснование затрат на создание и модернизацию системы защиты персональных данных требует применения формализованных методик оценки экономической эффективности и расчета возврата инвестиций в безопасность. Прямой расчет окупаемости в сфере информационной безопасности затруднен, поэтому

основной акцент делается на оценке предотвращенного ущерба.

Для оценки состояния защищенности применяются качественные и количественные метрики, такие как среднее время обнаружения инцидента, среднее время реагирования и локализации, процент успешно отраженных атак, а также показатель уровня охвата конечных узлов средствами мониторинга. Регулярный расчет и анализ данных метрик позволяет руководству организации принимать обоснованные решения по распределению бюджета, своевременно модернизировать устаревшие средства защиты и обеспечивать оптимальный баланс между безопасностью и удобством бизнес-процессов. **Заключение**. Защита персональных данных в современных информационных системах представляет собой многогранную задачу, требующую гармоничного сочетания правовых, организационных и инженерно-технических мероприятий. Успешное функционирование системы защиты невозможно без непрерывной адаптации к меняющемуся ландшафту угроз, постоянного совершенствования нормативной базы и повышения уровня культуры информационной безопасности среди сотрудников организации. Реализация предложенных в работе комплексных архитектурных решений и интеграция современных средств защиты позволяет обеспечить надежную сохранность персональных данных, минимизировать риски утечек и гарантировать соответствие самым строгим регуляторным требованиям.

Список литературы

1. Баранов А.А., Васильев В.П. Защита информации в распределенных информационных системах. — М.: Горячая линия - Телеком, 2022. — 356 с.

2. Григорьев С.М. Оценка рисков утечки персональных данных в корпоративных сетях // Вопросы защиты информации. — 2023. — № 2. — С. 14–22.

3. Дмитриев К.В., Елисеев И.О. Правовое регулирование защиты персональных данных: зарубежный и отечественный опыт // Информационное право. — 2021. — № 4. — С. 28–34.

4. Зайцев А.П., Шелупанов А.А. Технические средства защиты информации. — М.: Академия, 2020. — 448 с.

5. Иванов С.А., Кузнецов Н.А. Архитектурные подходы к построению защищенных облачных ИСПДн // Прикладная информатика. — 2023. — Т. 18, № 1. — С. 45–58.

6. Петров В.В. Моделирование угроз безопасности персональных данных в современных СУБД // Безопасность информационных технологий. — 2022. — № 3. — С. 89–97.

7. Сидоров Д.А. Применение концепции Zero Trust в системах защиты персональных данных // Защита информации. Inside. — 2023. — № 5. — С. 60–67.

8. Федоров Е.Н. Автоматизация аудита безопасности информационных систем персональных данных // Системы управления и информационные технологии. — 2021. — № 2. — С. 33–40.

9. Соколов М.Ю. Эффективность применения DLP-систем для предотвращения утечек ПДн // Безопасность данных. — 2022. — № 4. — С. 112–119.

10. Николаев А.С., Орлов П.Т. Криптографические методы защиты информации в открытых сетях. — СПб.: БХВ-Петербург, 2023. — 288 с.

Бет-әлпет биометриялық шаблондарын таралған сақтауға арналған құпияны табалдырықты бөлу схемаларын салыстырмалы талдау

Капалова Н.А.[1,2], Ергеш Н.С.[1,2]
e-mail: nkapalova@mail.ru, yergesh.nursultan@almatyedu.kz
[1]*ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институты, Қазақстан*
[2]*Әл-Фараби атындағы Қазақ ұлттық университеті, Қазақстан*

Аңдатпа. *Бет-әлпет биометриялық шаблоны -қайта шығарылмайтын тұрақты идентификатор, сондықтан оны бір түйінде бір көшірме түрінде сақтау сыни ақаулық нүктесін құрайды. Баяндамада авторлардың бет-әлпет шаблондарын құпияны табалдырықты бөлу арқылы таралған сақтау бойынша зерттеулерінің нәтижелері бірыңғай қорғау конвейері аясында жүйеленіп, үш схема салыстырылады: позициялық емес полиномдық санау жүйесі мен қалдықтар туралы қытай теоремасына негізделген схема (A), үлестерді тексеретін көп айнымалылы функциялар схемасы (B) және оның бет-әлпет биометриясына бейімделген блоктық-оралған модификациясы (C). Конвейер BioHashing түрлендіруін, аутентификацияланған шифрлауды және табалдырықты бөлуді біріктіреді; қорғау қабаты шығынсыз болғандықтан, LFW хаттамасындағы тең қателік деңгейі EER = 2,12 % ± 0,57 % өзгеріссіз сақталады. Барлық схемалар кворум бойынша жазбаны қатесіз қалпына келтіреді және бір түйіннің бұзылуына төзімді, бірақ табалдырықтан төмен ақпарат ағуы, үлестерді тексеру түрі, үлес көлемі мен жылдамдығы бойынша ерекшеленеді. Осы айырмашылықтар негізінде қолданылу ортасына (edge/IoT түйіндері, жоғары сенімділік талап ететін жүйелер, ірі шаблондар) қарай схеманы таңдау ұсыныстары тұжырымдалады.*

Түйін сөздер: *биометриялық шаблонды қорғау; бет-әлпет биометриясы; құпияны табалдырықты бөлу; позициялық емес полиномдық санау жүйесі; қалдықтар туралы қытай теоремасы; үлестерді тексеру; BioHashing.*

Кіріспе

Бетті тану биометриялық аутентификацияның ең кең таралған тәсілдерінің біріне айналды. Алайда пароль әшкереленгеннен кейін қайта шығарылады, ал бет-әлпет шаблоны адам өзгерте алмайтын физикалық белгіден алынады. Сондықтан сақталған шаблонның әшкереленуі қайтымсыз салдарға әкеледі: тұлғаны шектеусіз уақыт бойы жалған ұсынуға және оны тәуелсіз жүйелер арасында байланыстыруға мүмкіндік туады [1, 2].

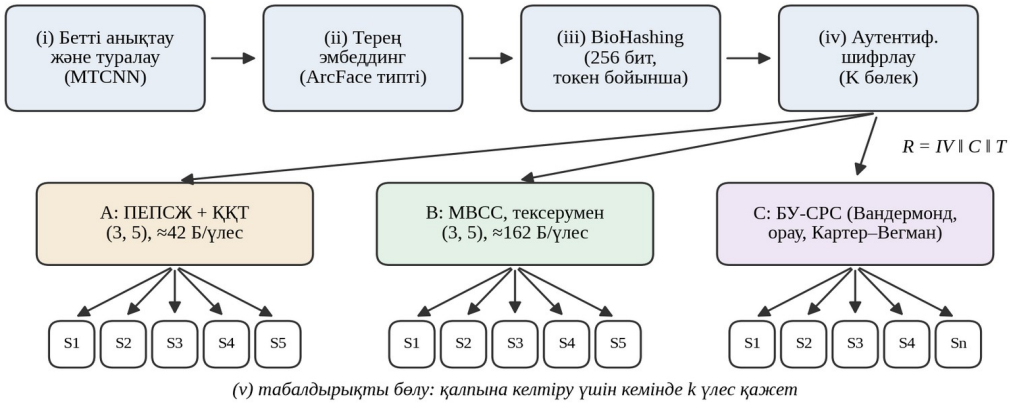
Практикада жиі кездесетін әлсіздік -қорғалған шаблонның бір ғана көшірмесінің бір ғана түйінде сақталуы. Шифрлау қолданылған жағдайда да мұндай сәулет барлық тәуекелді бір түйін мен бір кілтке шоғырландырады. Құпияны табалдырықты бөлу бұл мәселені тікелей шешеді: құпия n үлеске бөлінеді, кез келген k үлес оны қалпына келтіреді, ал k-дан аз үлес идеал жағдайда қарсыласқа пайдасыз [3, 4]. Дегенмен нақты схемалар табалдырықтан төмен ақпарат ағуы, үлестердің тұтастығын тексеру мүмкіндігі, үлес көлемі мен жылдамдығы бойынша айтарлықтай ерекшеленеді, және дәл осы айырмашылықтар белгілі бір қолданылу ортасына қай схема сәйкес келетінін анықтайды.

Баяндама авторлардың биометриялық деректерді кеңістікте таралған қоймада қауіпсіз сақтау бойынша зерттеулерінің жалғасы болып табылады. Бұған дейін саусақ ізі деректері үшін позициялық емес полиномдық санау жүйесіне (ПЕПСЖ) негізделген модификацияланған табалдырықты схема [5] және жалпы құпия деректер үшін үлестерді тексеретін көп айнымалылы функциялар схемасы (МВСС) [6] ұсынылған. [7] жұмыста бұл екі схема алғаш рет бет-әлпет шаблондарына қолданылып, ортақ платформада салыстырылды, ал МВСС схемасы бет-әлпет биометриясына бейімделген блоктық-оралған модификацияға (БУ-СРС) дейін дамытылды. Баяндаманың мақсаты -осы үш

схеманы бірыңғай қорғау конвейері аясында жүйелеу, олардың қасиеттерін ортақ критерийлер бойынша салыстыру және қолданылу ортасына қарай схеманы таңдау ұсыныстарын тұжырымдау.

1. Бет-әлпет шаблонын қорғау конвейері

Қорғау конвейері бет суретін тәуелсіз сақтау түйіндеріне таратылатын үлестер жиынына түрлендіреді және бес кезеңнен тұрады (сурет 1): (i) бетті анықтау және бес нүкте бойынша туралау (MTCNN [8]); (ii) терең эмбеддинг -L2-нормаланған нақты векторды алу (ArcFace типті кодтаушы [9]); (iii) қайтымсыз BioHashing түрлендіруі [10]; (iv) аутентификацияланған шифрлау; (v) қорғалған жазбаны табалдырықты бөлу. Верификация кезінде кворум жиналып, жазба қалпына келтіріледі, шифрдан шығарылады және жаңадан есептелген зонд-шаблонмен салыстырылады. Салыстырылатын A, B, C схемалары тек (v) кезеңде ерекшеленеді, бұл салыстыруды бақыланатын етеді.



Сурет 1. Бет-әлпет шаблонын қорғау конвейері: (i)–(iv) кезеңдер ортақ, (v) кезеңде A, B, C схемалары салыстырылады

BioHashing пайдаланушыға тән токеннен туындаған $\{r_i\}$ проекция бағыттары арқылы x эмбеддингін m-биттік бинарлық шаблонға түрлендіреді:

$$b_i = 1, \text{ егер } r_i^T x > \tau_i; \text{ әйтпесе } b_i = 0, \quad i = 1, \ldots, m, \tag{1}$$

мұндағы τ_i -табалдырықтар, $m = 256$. Түрлендіру қайтымсыз: таңба үлгісі проекция шамаларын жояды, ал токенді ауыстыру жаңа, байланыспайтын шаблон береді. Бинарлық шаблон кездейсоқ K кілтімен аутентификацияланған шифрланады:

$$(C, T) = AE_K(b), \quad R = IV \,\|\, C \,\|\, T, \tag{2}$$

мұндағы IV -инициализация векторы, C -шифрмәтін, T -аутентификация тегі (SHAKE-256 және BLAKE2b негізіндегі «алдымен шифрлау, кейін аутентификациялау» құрылымы, 128-биттік қауіпсіздік деңгейі). Маңызды жайт: кейін бөлінетін объект - биометриялық шаблонның өзі емес, 60 байттық R жазбасы, ал K кілті ешбір үлеске орналастырылмай, бөлек кілттерді басқару элементінде сақталады.

2. Салыстырылатын табалдырықты схемалар

2.1. A схемасы: ПЕПСЖ және қалдықтар туралы қытай теоремасы (ҚҚТ) [5, 7]. R жазбасының әрбір 16-биттік v блогы GF(2) өрісі үстіндегі $S(x) = x^{16} + \Sigma\, v_j x^j$ ($j = 0, \ldots, 15$) көпмүшесімен кодталады; x^{16} маркер мүшесі deg $S = 16$ болуын қамтамасыз етеді. Негіз ретінде алтыншы дәрежелі, жұптай өзара жай $n = 5$ келтірілмейтін $p_1(x), \ldots, p_5(x)$ көпмүшелері (бүтін санмен кодтағанда 67, 87, 103, 109, 115) алынады. i-ші үлес -қалдық:

$$a_i(x) = S(x) \bmod p_i(x), \quad i = 1, \ldots, n. \tag{3}$$

Кез келген $k = 3$ үлес дәрежесі $18 \geq 17$ болатын модуль береді, сондықтан $S(x)$ ҚҚТ бойынша бірмәнді қалпына келтіріледі:

$$S(x) = \left(\Sigma_{i \in I}\, a_i(x)\, M_{I,i}(x)\, N_i(x) \right) \bmod M_I(x), \tag{4}$$

мұндағы $M_I = \prod_{i \in I} p_i$, $M_{I,i} = M_I / p_i$, $N_i = M_{I,i}^{-1} \bmod p_i$. Екі үлес тек 12-дәрежелі модуль береді және $2^{17-12-1} = 16$ үміткер блок қалдырады. Мысалы, 12345 блогы үшін үлестер 37,

9, 22, 1 және 5 болады, ал олардың кез келген үшеуі 12345 мәнін қайтарады. Алтыншы дәрежелі келтірілмейтін көпмүшелер тек тоғыз болғандықтан, схема $n \le 9$ үлесті қолдайды. Бір үлес (қалдық, тұтастық тегі және идентификаторлар) шамамен 42 байт алады.

2.2. B схемасы: үлестерді тексеретін көп айнымалылы функциялар [6, 7]. R жазбасы $q = 2^{521} - 1$ жай өрісі үстіндегі $X = (x_1, \ldots, x_k)$ векторы ретінде кодталады, ал i-ші үлес -жария координаталық A_i векторы мен X векторының скаляр көбейтіндісі:

$$y_i = A_i \cdot X \bmod q = \left(\sum_j a_{ij} x_j \right) \bmod q, \quad i = 1, \ldots, n. \tag{5}$$

A_i коэффициенттері дилер тұқымынан SHAKE-256 арқылы детерминделген түрде алынады. Әр үлеспен бірге $UH = H(A_i)$, $signUH = H(UH \| SK_i)$ және $SH = H(y_i \| SK_i)$ SHA-256 тегтері сақталады, ал SK_i кілті тек верификаторда болады. Қалпына келтіру тегтерді тексергеннен кейін $k \times k$ сызықтық жүйені Гаусс әдісімен шешу арқылы орындалады. $t < k$ үлес q^{k-t} үміткер векторды қалдырады (тексерілетін рамп-схема), ал жалғыз құпия кездейсоқ толықтырумен қорғалғанда схема ақпараттық-теориялық мінсіз құпиялылыққа жетеді. Үлес көлемі -162 байт.

2.3. C схемасы: блоктық-оралған модификация БУ-СРС. Бұл схемада B схемасының ережесі үш бағытта өзгертілген: псевдокездейсоқ коэффициенттер детерминделген вандермонд ережесімен алмастырылған, жазбаны блоктық орау мен маскалау енгізілген, хеш-тексеру Картер–Вегманның бір реттік кодымен [11] алмастырылған. Жазба $q = 2^{127} - 1$ өрісінің L элементіне бөлінеді, $M = s_1 \| \ldots \| s_L$, және k -1 дәрежелі $P(x)$ көпмүшесіне оралады, мұндағы $k = t + L$:

$$P(\beta_m) = s_m, \ m = 1, \ldots, L; \quad P(\gamma_r) = \rho_r, \ r = 1, \ldots, t, \tag{6}$$

мұндағы β_m, γ_r -жұптай әртүрлі нөлдік емес нүктелер, ρ_r -біркелкі үлестірілген кездейсоқ маскалар. Түйін үлесі $y_i = P(x_i) \bmod q$, блоктар кез келген k үлестен Лагранж интерполяциясымен алынады, ал тұтастық тегі бір реттік (α_i, δ_i) кілтімен есептеледі:

$$\tau_i = \alpha_i y_i + \delta_i \bmod q. \tag{7}$$

Схеманың төрт қасиеті дәлелденген: кез келген k үлестен MDS-қалпына келтіру (вандермонд анықтауышы нөлге тең емес); t үлеске дейінгі мінсіз құпиялылық; $t \le m \le k$ үлес үшін ашылатын ақпараттың дәл бағасы $I = (m - t) \cdot \log_2 q$ бит; жалған үлестің тексеруден өту ықтималдығы $1/q$-дан аспайтын шартсыз тұрақтылық.

3. Эксперимент нәтижелері

Үш схема да, конвейердің ортақ кезеңдері де бірыңғай Python 3.12.3 платформасында іске асырылған; бастапқы код, бекітілген seed мәні және сандық нәтижелер файлдары ашық репозиторийде жарияланған [13]. Тану дәлдігі LFW верификация хаттамасында (View 2; 10 фолдтағы 6000 жұп) өлшенді (кесте 1). Қорғау қабаты шығынсыз (қалпына келтірілген шаблон тікелей есептелген шаблонмен бит-бит сәйкес), сондықтан BioHashing-тен кейінгі және толық конвейерден кейінгі EER мәндері бірдей.

Кесте 1. LFW хаттамасындағы тану дәлдігі (фолд бойынша орташа ± SD, 95 % сенімділік аралығы)

Баптау	Метрика	Шекті мән	EER	95 % CA
Бастапқы эмбеддинг	косинус	$\tau = 0{,}375$	1,37 % ± 0,62 %	[0,98; 1,75] %
BioHashing-тен кейін	Хэмминг	$d = 0{,}371$	2,12 % ± 0,57 %	[1,77; 2,47] %
Толық конвейерден кейін	Хэмминг	$d = 0{,}371$	2,12 % ± 0,57 %	[1,77; 2,47] %

EER-дің 0,75 пайыздық тармаққа өсуі тек BioHashing бинаризациясына байланысты. ISO/IEC 24745 [12] ұсынымдары бойынша есептелген байланыспаушылықтың жаһандық өлшемі $D_{sys} = 0{,}0615$ құрады (40 тұлға, әрқайсысы төрт тәуелсіз токенмен): әртүрлі токендермен тіркелген жазбаларды бір адамға тиесілі деп ажырату мүмкін емес. Схемалардың қасиеттері мен өлшенген көрсеткіштері кесте 2-де жинақталған.

Кесте 2. A, B және C схемаларын салыстыру

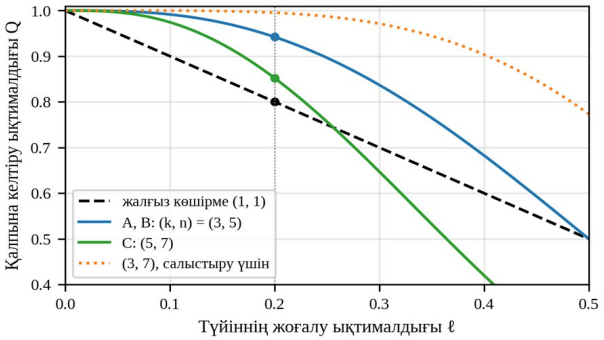
Қасиет	A: ПЕПСЖ–ҚҚТ [5, 7]	B: MBCC [6, 7]	C: БУ-СРС
Алгебралық негіз	GF(2)[x] үстіндегі қалдықтар, ҚҚТ	GF(q), q = 2^{521} -1; Гаусс әдісі	Z_q, q = 2^{127} -1; Лагранж интерполяциясы
Параметрлер	(k, n) = (3, 5)	(k, n) = (3, 5)	n = 7, k = 5, t = 2, L = 3
Қалпына келтіру кепілдігі	детерминделген (ҚҚТ)	ықтималдықтық (кездейсоқ коэффициенттер)	дәлелденген MDS (вандермонд)
Табалдырықтан төмен ағу	рамп: 16-биттік блоктан 6 / 12 бит (тек шифрмәтін)	тексерілетін рамп, q^{k-t} үміткер; жалғыз құпия режимінде мінсіз	t-ге дейін мінсіз; t ≤ m ≤ k үшін (m -t)·$\log_2 q$ бит
Үлесті тексеру	4 Б SHA-256 тегі (2^{-32})	SHA-256 тегтері (2^{-256}, есептеу тұрақтылығы)	Картер–Вегман коды (≤ 1/q, шартсыз)
Үлес көлемі	≈ 42 Б (60 Б жазба)	≈ 162 Б (60 Б жазба)	256 Б жазбада байттық бөлуден 64 есе аз
Бөлу / қалпына келтіру, мс	0,34 / 1,49	0,075 / 2,65	0,588 / 0,359*
Шабуылдар бойынша тексеріс	бұрмалау 2000/2000 анықталды; жалған үлес 0/20 000	бұрмалау 2000/2000 анықталды; жалған үлес 0/20 000	жалған үлес 0/200 000; қалпына келтіру 6300/6300

* C схемасының уақыты 256 байттық жазба мен $q = 2^{127}$ -1 өрісі үшін өлшенген, сондықтан A және B бағандарымен тікелей салыстыруға келмейді.

Түйіндердің жоғалуына төзімділік. Әр түйін ℓ ықтималдықпен тәуелсіз жоғалады деп есептегенде, кемінде *k* үлестің сақталу ықтималдығы:

$$Q(n, k, \ell) = \Sigma_{i=k}^{n} C(n, i)(1 - \ell)^i \ell^{n-i}. \quad (8)$$

ℓ = 0,2 болғанда (3, 5) конфигурациясы 94,2 %, (5, 7) конфигурациясы 85,2 % жағдайда жазбаны қалпына келтіреді, ал жалғыз көшірмені сақтау тек 80 % жағдайда сәтті болады (сурет 2). Табалдырықты арттыру құпиялылықты күшейтеді, бірақ қолжетімділікті төмендетеді, сондықтан *k* мен *n* бірлесіп таңдалуы керек.



Сурет 2. Түйіндер тәуелсіз жоғалған кездегі қорғалған шаблонның қолжетімділігі (8-формула)

Шабуылдарға төзімділік. A және B схемаларында ұрланған токен мен табалдырықтан аз үлестен шаблонды қалпына келтіру және BioHashing-ті кері айналдыру 2000 сынақтың бірде-бірінде сәтті болмады [7]. Z_{251} шағын өрісінде кездейсоқ коэффициенттері бар бастапқы схемада 3-тен-5 ішкі жиындардың 0,42 %-ы азғындалған болса, C схемасының вандермонд ережесінде бірде-біреуі азғындалмады.

4. Талқылау және схеманы таңдау ұсыныстары

Қалпына келтірудің дұрыстығы, тану дәлдігі және түйіндердің жоғалуына төзімділік бойынша үш схема эквивалентті және жалғыз көшірмені сақтаудан басым. Оларды ажырататын осьтер -табалдырықтан төмен ағу, үлестерді тексеру түрі және ресурс шығыны. A схемасы ықшам (≈ 42 Б), детерминделген және тез қалпына

келтіріледі, сондықтан шектеулі ресурстары бар edge және IoT түйіндері үшін табиғи таңдау болып табылады; оның бағасы -рамп-ағу: бір үлес 16-биттік блоктың 6 битін, екі үлес 12 битін ашады. Бұл биттер биометрияның емес, шифрмәтіннің биттері екенін атап өткен жөн: *K* кілті бөлек сақталғанда және BioHashing қайтымсыз болғанда биометриялық ақпараттың ағуы нөлге тең. B схемасы үлестерді табиғи түрде тексеруді және жалғыз құпия режимінде мінсіз құпиялылықты ұсынады, бірақ үлестері үлкенірек, қалпына келтіруі баяуырақ; ол жоғары сенімділік талап етілетін жүйелерге сәйкес келеді. C схемасы B схемасының ықтималдықтық қалпына келтіру кепілдігін дәлелденген MDS қасиетіне, есептеу тұрақтылығы бар тексеруді шартсыз тексеруге ауыстырады және 256 байттық жазба үшін үлестердің жалпы көлемін байттық бөлумен салыстырғанда 64 есе (61 440 Б-тан 960 Б-қа дейін) қысқартады; оның бағасы -*L* параметрімен басқарылатын рамп-аралық ($t < m < k$). Сондықтан C схемасы ірі шаблондар мен ұзақ мерзімді таралған қоймалар үшін ұсынылады, ал $L = 1$ болғанда ол толық мінсіз схемаға айналады.

Барлық қауіпсіздік тұжырымдары қатерлер моделіне шартты: *K* кілті мен BioHashing токені үлестерден бөлек сақталады, аутентификацияланған шифр бұзылмаған деп есептеледі; бұл шарт бұзылса, A схемасы үшін дәлел әлсірейді және B немесе C схемаларына басымдық беру керек. Тану дәлдігі фронтенд кодтаушысының қасиеті және басқа кодтаушыда өзгеруі мүмкін.

Қорытынды

Баяндамада бет-әлпет биометриялық шаблондарын таралған сақтауға арналған үш табалдырықты схема -ПЕПСЖ мен ҚҚТ негізіндегі схема, үлестерді тексеретін көп айнымалылы схема және оның блоктық-оралған модификациясы -BioHashing пен аутентификацияланған шифрлауды қамтитын бірыңғай конвейер аясында жүйеленіп, салыстырылды. Барлық схемалар жазбаны кворум бойынша бит-бит қалпына келтіреді, тану дәлдігін өзгертпейді және бір түйіннің бұзылуына төзімді. Схемалардың айырмашылығы сандық түрде бағаланған ымыраға келтіріледі: A -edge/IoT үшін ықшам рамп-схема, B -тексерілетін және жалғыз құпия үшін мінсіз схема, C -дәлелденген MDS, шартсыз тексеру және көлемді үнемдеу қасиеттерін біріктіретін схема. Дұрыстық дәлелдері тек жазбаның байттық құрылымына тәуелді болғандықтан, сәулет басқа биометриялық модальдылықтарға өзгеріссіз қолданылады. Алдағы жұмыстар үш схеманы бірдей жазбада бірыңғай стендте тікелей өлшеуге және үлестерді динамикалық жаңартуға бағытталады.

Қаржыландыру. Жұмыс ҚР Ғылым және жоғары білім министрлігі Ғылым комитетінің AP23487787 «Кеңістікте таралған қоймада аутентификация жүйелерінің биометриялық деректерін қауіпсіз сақтау алгоритмін әзірлеу» гранты аясында орындалды.

Әдебиеттер

1. Jain, A. K. Biometric template security / A. K. Jain, K. Nandakumar, A. Nagar // EURASIP Journal on Advances in Signal Processing. -2008. -Vol. 2008. -Art. 579416.

2. Nandakumar, K. Biometric template protection: bridging the performance gap between theory and practice / K. Nandakumar, A. K. Jain // IEEE Signal Processing Magazine. -2015. -Vol. 32, N. 5. -P. 88–100.

3. Shamir, A. How to share a secret / A. Shamir // Communications of the ACM. -1979. -Vol. 22, N. 11. -P. 612–613.

4. Blakley, G. R. Safeguarding cryptographic keys / G. R. Blakley // Proceedings of the AFIPS National Computer Conference. -New York, 1979. -Vol. 48. -P. 313–317.

5. Nyssanbayeva, S. Modified Shamir threshold scheme for secure storage of biometric data / S. Nyssanbayeva, N. Kapalova, S. Beisenova // Computers. -2026. -Vol. 15, N. 4. -Art. 228.

6. Kapalova, N. Secret sharing scheme with share verification capability / N. Kapalova, A. Haumen, K. Algazy // Computers. -2025. -Vol. 14, N. 9. -Art. 393.

7. Kapalova, N. Protecting facial biometric templates with threshold secret sharing: a comparative resource-aware study of a non-positional polynomial scheme and a multivariable verification scheme / N. Kapalova, N. Yergesh // Computers. -2026. -Vol. 15, N. 9. -Art. 604.

8. Zhang, K. Joint face detection and alignment using multitask cascaded convolutional networks / K. Zhang, Z. Zhang, Z. Li, Y. Qiao // IEEE Signal Processing Letters. -2016. -Vol. 23, N. 10. -P. 1499–1503.

9. Deng, J. ArcFace: additive angular margin loss for deep face recognition / J. Deng, J. Guo, N. Xue, S. Zafeiriou // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. -Long Beach, 2019. -P. 4690–4699.

10. Teoh, A. B. J. BioHashing: two factor authentication featuring fingerprint data and tokenised random number / A. B. J. Teoh, D. C. L. Ngo, A. Goh // Pattern Recognition. -2004. -Vol. 37, N. 11. -P. 2245–2255.

11. Wegman, M. N. New hash functions and their use in authentication and set equality / M. N. Wegman, J. L. Carter // Journal of Computer and System Sciences. -1981. -Vol. 22, N. 3. -P. 265–279.

12. ISO/IEC 24745:2022. Information security, cybersecurity and privacy protection - Biometric information protection. -Geneva: ISO, 2022.

13. Yergesh, N. FACEID: facial biometric template protection platform (source code and results). -https://github.com/yergeshnursultan-dev/FACEID (қаралған күні: 08.08.2026).